

**THE IDENTIFICATION AND CHARACTERIZATION OF COPY NUMBER VARIANTS
IN THE BOVINE GENOME**

A Dissertation

by

Ryan N Doan

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Scott Dindot
Committee Members,	Noah Cohen
	William Murphy
	Loren Skow
	James Womack
Intercollegiate Faculty Chair,	Craig J. Coates

August 2013

Major Subject: Genetics

Copyright 2013 Ryan N Doan

ABSTRACT

Separate domestication events and strong selective pressures have created diverse phenotypes among existing cattle populations; however, the genetic determinants underlying most phenotypes are currently unknown. *Bos taurus taurus* (*Bos taurus*) and *Bos taurus indicus* (*Bos indicus*) cattle are subspecies of domesticated cattle that are characterized by unique morphological and metabolic traits. Because of their divergence, they are ideal model systems to understand the genetic basis of phenotypic variation. Here, we developed DNA and structural variant maps of cattle genomes representing the *Bos taurus* and *Bos indicus* breeds. Using this data, we identified genes under selection and biological processes enriched with functional coding variants between the two subspecies. Furthermore, we examined genetic variation at functional non-coding regions, which were identified through epigenetic profiling of indicative histone- and DNA-methylation modifications. Copy number variants, which were frequently not imputed by flanking or tagged SNPs, represented the largest source of genetic divergence between the subspecies, with almost half of the variants present at coding regions. We identified a number of divergent genes and biological processes between *Bos taurus* and *Bos indicus* cattle; however, the extent of functional coding variation was relatively small compared to that of functional non-coding variation. Collectively, our findings suggest that copy number and functional non-coding variants may play an important role in regulating phenotypic variation among cattle breeds and subspecies.

DEDICATION

I dedicate this to my wife, Ashley, for all of her help and support. By always being there when I need her most, she is the most important reason for my successes in life. I also dedicate this to my parents, Neil and Julie, and brother, Chris, who have always supported me throughout my life. Their strong support allowed me to become the person I am today.

ACKNOWLEDGEMENTS

I would like to thank Dr. Scott Dindot for his support and guidance throughout the past four years. The experiences gained during my research will play a key role in my future endeavors. I would also like to thank all of my committee members, Drs. Noah Cohen, William Murphy, Loren Skow, and James Womack for all of their help and guidance during my dissertation project.

I would also like to thank all of the people who have been involved in this project who have provided intellectual guidance, access to samples and assistance with experiments.

TABLE OF CONTENTS

	Page
ABSTRACT	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS.....	v
LIST OF FIGURES	ix
LIST OF TABLES	xii
CHAPTER I INTRODUCTION	1
Types of Genetic Variation among Individuals.....	1
Mechanism of Structural Variant Formation	3
Mutation Rates of Structural Variants.....	7
Methods to Identify Structural Variation.....	7
Detection of Copy Number Variants Using SNP Arrays.....	9
Whole-Genome Sequencing	11
Epigenetic Gene Regulation.....	13
CNVs and Trait Association	15
Bovine Evolution	16
Importance of Understanding Traits in Cattle	17
Bovine Genomics.....	19
Use of Next-Generation Sequencing to Examine Genetic Variation in Cattle ...	21
The Need for a Centralized Variant Database.....	22
Dissertation Strategy for Characterizing CNVs In The Bovine Genome	23
CHAPTER II CGH ANALYSIS OF CNVS.....	25
Introduction	25
Methods	28
DNA Sample	28
Comparative Genomic Hybridization Array Designs	28
Array Comparative Genomic Hybridization (aCGH) Methods	29
CNV Confirmation	31
Genomic Characterization of CNV Content	32

Analysis of CNV Content in the Bovine Umd3.1 Assembly	33
Functional Analysis of CNVs	34
Population Analysis of CNVs	35
Results	36
CGH Array Design.....	36
Exome CGH Accuracy and Resolution	40
Genome Distribution of CNVs	45
Umd3.1 Analysis	50
Functional Analysis	52
Population Analysis.....	57
Comparative Analysis.....	59
Discussion	61
 CHAPTER III WHOLE GENOME ANALYSIS OF <i>Bos taurus</i> AND <i>Bos indicus</i> COWS.....	 63
Introduction	63
Methods	65
Whole-Genome Sequencing	65
Sequence Mapping	66
Variant Detection.....	66
Variant Confirmation.....	68
Genetic Variant Annotation and Analysis.....	71
Genotyping for Known Mutations.....	72
Evolutionary Analysis	72
Results	72
Genome Sequencing.....	72
Variant Identification and Annotation	74
Functional Analysis of Variants	83
Genotyping for Known Mutations.....	85
Evolutionary Analysis	88
Discussion	90
 CHAPTER IV VARIATION AT REGULATORY ELEMENTS.....	 95
Introduction	95
Methods	97
Epigenetic Profiling.....	97
Results	99
Identification of Regulatory Elements in the Bovine Genome	99
Characterization of Variant Densities	105
Discussion	106

CHAPTER V	CNV IMPUTATION	109
	Introduction	109
	Methods	111
	BovineHD BeadChip CNV Analysis.....	111
	PLINK Analysis	112
	CNV Imputation by SNPs	112
	Results	115
	Genotype Analyses	115
	CNV Analysis	116
	CNV Imputation by SNP Genotyping Array	117
	Discussion	119
CHAPTER VI	CATHELICIDIN ANALYSIS.....	121
	Introduction	121
	Methods	123
	Population Analysis of the Cathelicidin Duplication.....	123
	RNA Isolation	124
	Expression Analysis	125
	Induction of Cathelicidin Expression in Fibroblasts.....	126
	Results	127
	Population Analysis of Cathelicidin CNV	127
	Expression Profile of <i>CATHL1</i> and <i>CATHL4</i>	129
	Association <i>CATHL1</i> and <i>CATHL4</i> Duplication with Expression.....	130
	Discussion	131
CHAPTER VII	CONCLUSIONS	134
REFERENCES	141
APPENDIX 2.1.....		162
APPENDIX 2.2.....		164
APPENDIX 2.3.....		191
APPENDIX 2.4.....		196
APPENDIX 2.5.....		197
APPENDIX 2.6.....		198
APPENDIX 2.7.....		199

APPENDIX 2.8.....	200
APPENDIX 2.9.....	203
APPENDIX 2.10.....	204
APPENDIX 2.11.....	208
APPENDIX 2.12.....	211
APPENDIX 3.1.....	217
APPENDIX 3.2.....	223
APPENDIX 3.3.....	224
APPENDIX 3.4.....	225
APPENDIX 3.5.....	226
APPENDIX 4.1.....	227
APPENDIX 4.2.....	228
APPENDIX 4.3.....	229
APPENDIX 5.1.....	230
APPENDIX 5.2.....	231
APPENDIX 5.3.....	232

LIST OF FIGURES

FIGURE	Page
1.1 Genomic and Epigenomic Analysis Pipeline.....	24
2.1. Gene Coverage of Bovine Exome CGH Array	37
2.2 Samples Used on Exome CGH Analysis	41
2.3 Identification and Sanger Sequencing Confirmation of a 98bp Loss within UTR of <i>CYYR1</i> in Nellore-1	43
2.4 Confirmation of CNV Used as Smallest Size for CGH Analysis	43
2.5 Confirmation of 264bp Homozygous Deletion Affecting the Majority of a Single Exon.....	44
2.6 (A.) Identification and Confirmation through (B.) PCR and (C.) qPCR of a Large Complex CNV Region on Chromosome 5 of Angus and Nellore Cattle.....	45
2.7 Enrichment Plot of Exome CNVs Based on the Log ₁₀ of Variant Lengths	47
2.8 Plot of Percentage Exome CNVs Overlapping Segmental Duplications.....	48
2.9 Identification of a Complex CNV Flanking 5' Region of Gene Consisting of (A.) Flanking Loss and Gain, (B.) Gains and (C.) No CNV.....	50
2.10 Comparison of the Cathelicidin Locus in the (A.) Bostau4.0 and (B.) Umd3.1 Genome Assemblies	52
2.11 Functional Analysis of Genes Affected by CNVs in the Exome CGH Array.....	53
2.12 Exon Resolution Analysis of Single Exon CNV in <i>GDF9</i>	56
2.13 Population Structures of CNVs among Each Breed.....	58
2.14 Sharing of Genes Affected by CNVs among Diverse Species	60
3.1 Read-Depths across Assembled Chromosomes	74

3.2	Distribution of Indels by Variant Length	80
3.3	Biological Process (BP) Analysis of SNVs, Indels, and CNVs.....	84
3.4	Biological Process (BP) Analysis of Radical and Conservative nsSNVs.....	85
4.1	Definition of Genomic Regions Used To Identify Regions (B.) With H3k4me3, (C.) DNA Methylation and (D.) With Differential Modifications	100
4.2	(A.) SNVs Located Within Genomic Regions (B.) With H3k4me3, (C.) DNA Methylation and (D.) With Differential Modifications	101
4.3	Identification of Gene Known to Undergo X-Inactivation (<i>MAOA</i>) By Overlapping Histone and DNA Methylation	104
4.4	Identification of a Gene Known to Escape X-Inactivation In Humans (<i>KDM6A</i>) By Histone and DNA Methylation Analysis In Cattle.....	104
4.5	Confirmation of Gene that is known to be Imprinted (<i>SNRPN</i>) By Overlapping Histone and DNA Methylation in Cattle.....	104
4.6	Potential Allelic Exclusion of <i>Bola</i> Gene Indicated By Overlapping Histone and DNA Methylation in Cattle.....	105
4.7	Comparison of SNVs at Genetic Elements	106
5.1	Diagram of Method Used To Identify CNVs Tagged By Homozygous SNPs within the Variant.....	113
5.2	Diagram of Method Used To Impute CNVs by Homozygous SNPs Flanking the Variants.....	114
5.3	Diagram of Complete CNV Imputation through the Combination of SNPs Within and Flanking CNVs.....	114
5.4	Hierarchical Clustering of Cattle Using BovineHD SNP Genotypes	116
5.5	Example of CNV that is Not Accurately Imputed by BovineHD SNP Array Genotypes	119
6.1	Population Analysis of Cathelicidin CNV in <i>Bos Indicus</i> and <i>Bos Taurus</i> Cattle.....	128
6.2	Expression Profiles of (A.) <i>CATHL1</i> and (B.) <i>CATHL4</i> Across	

Several Tissues from Angus Cattle.....	130
6.3 Expression Profile of (A.) <i>CATHL1</i> and (B.) <i>CATHL4</i> in White Blood Cells of Cattle with (Nellore) and Without (Angus) The Cathelicidin CNV	131

LIST OF TABLES

TABLE	Page
1.1 Trends in the Beef Cattle Industry	18
2.1 Statistics of Exome CGH Array Design.....	38
2.2 Comparison of Exome CGH and Nimblegen Tiling Arrays.....	39
2.3 False Discovery Rates Of CNVs by Classification	42
2.4 Identification of CNVs and Affected Genes across the Bovine Exome.....	46
2.5 Probe Statistics for CNV Tiling Array.....	49
2.6 Annotation of CNV Tiling Array Probes.....	49
2.7 CNV Genes with Associated OMIA Terms.....	54
3.1 Definitions of Terms Used For the Evaluation of Accuracy SNV Identification	69
3.2 Overview of Whole-Genome Sequencing Data	73
3.3 Evaluation of Accuracy of SNV Identification Using Relaxed Filtering.....	75
3.4 Annotation of SNVs Identified In Comparison To the Reference Hereford Genome.....	77
3.5 Annotation of SNVs Identified Between the Angus And Nellore Genomes	78
3.6 Annotation of Indels Identified In Comparison To The Reference Hereford.....	79
3.7 Annotation of Indels between the Angus and Nellore Genomes	80
3.8 Independent and Comparative CNV Analyses Using Sequence Read-Depth.....	82
3.9 Genotyping for Known Casual and Associated Mutations For Diseases and Traits in Cattle.....	86
3.10 Genes with Significant Positive Selection Using the Branch-	

Specific Model.....	89
3.11 Effects of Selection Using Homozygous SNVs within Refseq Genes.....	90
4.1 Summary of Epigenetic Profiling For H3k4me3 and DNA Methylation.....	100
5.1 SNP Genotype Distributions from BovineHD Array.....	115
5.2 Complete CNV Imputation Using Genotypic Data From The BovineHD SNP Array	118
6.1 Taqman PCR Primers for Genomic Analysis of CNV.....	124
6.2 Primers for Expression Analysis	125
6.3 Reaction Mixture for Expression Analysis.....	126

CHAPTER I

INTRODUCTION

TYPES OF GENETIC VARIATION AMONG INDIVIDUALS

The structure and composition of genomes has traditionally been considered to be constant with very few changes occurring, even among species divergent by millions of years. While the long held notion that any two individuals are more than 99.9% genetically identical may still apply, recent studies have clearly shown that our theory of genetic diversity needs some slight modifications [1-3]. Two individuals have very similar general chromosomal structures and coding sequence, but intra-chromosomal structural variation affects up to 10% of individual genomes.

The most well studied type of genetic variants are single nucleotide variants (SNVs) caused by mistakes during DNA replication. The characterization of SNVs in numerous genomes has identified millions of single nucleotide polymorphisms (i.e., SNVs with characterized allelic frequencies known as SNPs). An individual human has an average of 3.3 million SNPs [4-6]. These SNPs have the potential to alter critical bases in coding and regulatory regions of a genome, thereby causing dramatic phenotypic effects.

Insertion and deletion variants (INDELs) are another type of genetic variant caused by errors during DNA replication and retrotransposition of DNA elements. These variants are often single base insertions and deletions occurring by polymerase slippage. Additionally, INDELs can be formed through LINE/SINE insertions or microhomology mediated excision and duplication. While the definition of an INDEL has changed over time, it is generally accepted that they can range in lengths of a single base up to 50 base-pairs (bp) [1, 4]. The formation of an INDEL within a coding

portion of a gene has the potential to be much more deleterious to the gene's function. Insertion and deletion variants often lead to a frameshift resulting in the formation or loss of stop codons. The altered transcripts typically undergo nonsense-mediated decay, causing a reduction in functional protein in the cell. The strong negative selective pressure against deleterious INDELs resulted in their abundance within intergenic and non-coding regions.

Structural variants (SV) in a genome can be either balanced (no change in DNA content) or unbalanced (changes DNA content). Balanced structural rearrangements in a genome include inversions and translocations of a segment of DNA to another location in the genome. Unbalanced structural rearrangements (i.e., known as copy number variants [CNVs]), result in the duplication or deletion of a segment of DNA. Over the past decade, CNVs have been intensively studied in humans, revealing that as much as 10% of the human genome is affected by CNVs. The size of a CNV was traditionally considered a region greater than 1,000 bp (1 kb); however, with the increased resolution of detection methods, CNVs can be as small as 50 bp in length. The effects of CNVs are dependent on the size and region affected. Copy number variants encompassing entire genes can result in overall increases and decreases of gene expression. The occurrence of CNVs within single or several exons of a gene may lead to the fusion of two adjacently located genes, new exons, or nonsense mediated decay. Additionally, CNVs overlapping genic and intergenic regulatory elements can alter splicing and the level of gene expression [7, 8].

Ancestral duplications are shared within a population and have, over time, acquired mutations resulting in up to 10% sequence divergence between the duplicated regions. This class of unbalanced structural variant is known as a segmental duplication

(SD) and, while similar to a CNV, they differ in their sequence identity and ancestry. Segmental duplications are found throughout the genome but are enriched at regions within 1Mb of centromeres (pericentromeric) and subtelomeric regions. While segmental duplications are distinct from CNVs, they are often highly enriched for new CNVs due to the predisposition to recombination events from the sequence similarity.

Overall, the size and distribution of CNVs throughout genomes are highly dependent on predisposition due to the presence of repetitive elements and homologous regions. The combination of several mechanisms of formation leads to CNV enrichments in approximate size groups of 350-, 6,000-, and 50,000-bp. Regions containing sequence repeats, extended homology and simple structures predispose the regions to the formation of new and recurring structural variants. Additionally, regions of ancestral segmental duplications have been shown to contain 20-30% of CNVs in the human genome [9, 10]. Comparative analysis has revealed that approximately 27% of human CNVs are shared with at least one species of primate [10, 11]. Therefore, CNVs are both ancestrally inherited and novel within individuals.

MECHANISM OF STRUCTURAL VARIANT FORMATION

The structures of genomes are constantly undergoing alterations in size, content and organization. Large-scale changes in genome structure, known as structural variation, arise from numerous paths involving recombination and DNA replication. The type of variant, mechanism of formation, and prevalence within a population are highly dependent on the genetic makeup of the genomic region. The major contributor to the formation of structural variation is various types of repetitive content in the genome including short interspersed nuclear elements (SINEs), long interspersed elements

nuclear (LINEs), variable number tandem repeats (VNTRs), and long terminal repeats (LTRs).

Approximately 50% of the human genome is repetitive, including more than 500,000 LINEs and 140,000 SINEs [11]. The most prevalent LINE, L1, is enriched in AT-rich regions, while the Alu (SINE) is enriched within GC-rich regions [12-15]. Long terminal repeats have been shown to build up within intergenic regions. The activity of retrotransposons has drastically reduced over time, with LINEs being the only readily active element in the genome even though 99% remain inactive [11]. Also, the rates of element activity vary in different genomes, with the largest burst of activity occurring 30-40 million years ago during primate evolution [10, 11, 16].

Non-allelic homologous recombination (NAHR) is a crossing over event between 2 regions with high sequence similarity [17, 18]. Unequal crossing over often occurs during homologous recombination in prophase I of meiosis 1 in diploid cells. However, NAHR also occurs during mitosis in somatic cells, resulting in mosaicism for structural variation. Somatic NAHR occurs in numerous cancers, patients with neurofibromatosis, and healthy individuals [17, 19-22]. Non-allelic homologous recombination often results in large CNVs due to the size requirements for recombination to occur. The minimal efficient processing segments (MEPS) are the minimal length of genomic regions with high sequence similarity required for efficient NAHR. The length of MEPS in meiosis ranges from 300-500bp, while in mitosis it can be as short as 114bp [17, 22, 23]. The recombination between regions of sequence similarity can result in both balanced and unbalanced structural changes. The balanced structural changes can lead to unbalanced changes in future generations [24, 25]. The formation of ancestral unbalanced variants leads to segmental duplications that are shared within a

population. The repetitive nature of these regions makes them prone to acquiring further CNVs.

Additionally, the repair of double-strand breaks without the use of sequence homology results in translocations and telomere fusions through non-homologous end joining (NHEJ) [26, 27]. Non-homologous end joining proceeds through 4 main steps: detection of double-stranded break, bridging of broken ends, modification of ends, and ligation. Unlike NAHR, MEPS and repeats are not required for NHEJ. However, NHEJ results in an 'information scar' of several hundred bases at the breakpoint [17]. Errors occurring during this mechanism lead to deletion and duplication events.

The formation of SVs can also occur through replication mechanisms including microhomology-mediated break-induced replication (MMBIR) and fork stalling, and template switching (FoSTeS) [13, 14, 25, 27-30]. The proposed mechanisms of MMBIR and FoSTeS both involve errors at the replication forks. MMBIR involves the breakdown of a replication fork, causing a template switch by forming a new fork with another template [30]. Eventually the replication returns to the original sister chromatin and continues replication as normal, with a new segment of DNA. FoSTeS is similar to MMBIR, but the switches occur due to fork stalling not from a break in the DNA [17, 31].

Tandem repeats are highly polymorphic and can undergo expansions and contractions, leading to their variable numbers. VNTRs may undergo expansion and contraction using many methods, including strand-slippage recombination and recombination [28]. Strand-slippage recombination occurs when the newly synthesized strand denatures from the template and anneals to a different region of the array of repeats, resulting in looping of either the template for the new DNA strand. Depending on which strand undergoes looping, the repeat array can either expand or contract.

Alternatively, inter- and intra- repeat recombination can result in the expansion and contraction of VNTRs.

Retrotransposition of mobile elements in the genome is a major cause of structural variation in genomes. While the majority of mobile elements in a genome are inactive remnants of ancestral events, a few types of elements (e.g., L1, Alu, and SINE-VNTR-Alu elements) remain active. The insertion of mobile elements (i.e., mobile element insertion, MEI) results in both deletions and insertions in a genome, relative to a reference genome [12, 14]. Transposable elements can be classified into three groups based on their mechanism of transposition [32]. The first group, DNA transposons, contains inverted terminal repeats and a single open reading frame (ORF) encoding transposase. The transposase moves the DNA transposons through the genome using a 'cut and paste' mechanism lacking an RNA intermediate. Autonomous retrotransposons (e.g., LTRs [HERV] and Non-LTRs [L1]) move using a DNA-RNA-DNA process through the utilization of an ORF encoding proteins for retrotransposition (e.g., nucleic acid binding protein, endonuclease, and reverse transcriptase) [32]. Non-autonomous retrotransposons (e.g., Alu elements) rely on retrotransposition proteins from other elements (e.g., LINES) to move throughout a genome [33].

Collectively, the formation of small CNVs between 50 and 10,000bp in length are likely due to mobile element insertions (MEI), tandem repeats expansion/contraction and MMBIR. The enrichments of smaller CNVs (350bp, 6,000bp) have been attributed to insertions and deletions of Alu elements and LINEs, respectively [13, 34]. Mobile element insertion is considered a major cause of insertion SVs in the human genome [12, 35]. Also, the insertion of mobile elements into coding genes leads to novel exons, exonization and alternative splicing [36]. Retrotransposition is linked to approximately

30% of small insertions and deletions in the human genome [13, 37]. Additionally, the expansion and contraction of VNTR account for 2% of SVs in a genome [38]. The impact of VNTRs can be significant due to the fact that approximately 17% of coding gene possess repeats within coding exons [38].

MUTATION RATES OF STRUCTURAL VARIANTS

The occurrence of structural variants depends on several conditions such as length, sequence complexity and homology. The commonly cited rate for a single base change is in the order of 10^{-8} per base per generation [39, 40]. Regions containing tandem repeats, LINES and SINEs, have been shown to have typical mutation rates of 10^{-3} to 10^{-7} ; however, rates as high as 10^{-2} have been observed [28, 29, 41, 42]. The rates for CNV formation are highly dependent on the structure of the surrounding region and the length of the CNV, but are commonly believed to be approximately 10^{-4} [40, 43]. Regions containing CNVs and SD's increase the likelihood that another event will occur in the same region. Overall, there are very few *de novo* mutations giving rise to CNVs in a single genome [34, 43]. The majority of CNVs are inherited from the parents [44]. With mutation rates and variant lengths much higher than single base mutations, CNVs affect a larger percentage of the coding and non-coding portions of the genome than any other type of genetic variant.

METHODS TO IDENTIFY STRUCTURAL VARIATION

Historically, CNVs were investigated by cytogenetic methods (e.g., fluorescent *in situ* hybridization [FISH], G-banding, etc). Despite limitations in resolution for detecting CNVs, these methods were commonly utilized, until array-based methods were developed. Since then, the study of CNVs has been highly dependent on comparative genomic hybridization (CGH), which was developed in 1992 [45]. The methods have

greatly improved since procedures where chromosomal spreads are used as controls to which a labeled sample is hybridized. Soon after the creation of the CGH concept, focused arrays were developed using cosmids, p1-derived artificial chromosomes, yeast artificial chromosomes, and bacterial artificial chromosomes [45-50]. The resolution of CGH technology was further reduced by spotting cDNA and PCR fragments onto glass slides [46, 51, 52]. The newest CGH arrays can contain more than two million oligonucleotides (oligos) on a single 1 inch x 3 inch glass slide. Despite the advancements in hybridization and probe generation techniques, the general concept remains the same. The basic premise is that a control and test sample are differentially labeled, usually with cy3 and cy5 fluorophores, and then competitively hybridized to an array containing single-stranded oligos matching the region of interest. Once the samples are bound to the oligo spots on the glass slide, an image of the laser-excited samples is analyzed to compare the intensities of the cy3 and cy5 dyes at each probe. In general, if the control and test samples contain equal copies of the genomic region, the intensities will be the same; however, if they do not have equal copies, a difference in signal intensities will be observed.

Array CGH (aCGH) has become an essential tool in both the research and medical industries for the detection of CNVs in a genome. However, there are limitations to the array designs. A significant limitation is the placement of the probes: their need for unique sequence prevents the placement in a large portion of the genome. Also, since aCGH is a comparative analysis, CNVs shared between the reference and the control cannot be identified. Additionally, the comparative nature of aCGH makes it difficult to determine which sample actually has the gain or loss without performing additional investigation. Another limitation is the ability of most arrays to

detect small CNVs. Until recently, most arrays were unable to identify CNVs smaller than 1kb, which was due to the tiling density of probes within a region. Recent advancements in printing technology have allowed for the creation of dense tiling arrays and have reduced the minimum resolution to only a few hundred base pairs [53, 54]. A final limitation is platform bias, which is due to differences in chemistry used during the array manufacturing process [55]. Furthermore, there are different CNV calling requirements among the commercially available arrays. For example, an Agilent CGH array requires only 3 probes to be accurate while Nimblegen and Illumina arrays require 10 probes [56, 57].

DETECTION OF COPY NUMBER VARIANTS USING SNP ARRAYS

SNP arrays for genotyping and performing GWAS studies have been widely used since their creation more than a decade ago, with 1,518 human studies being published to date ([58], <http://www.genome.gov/gwastudies/>). Recently, the ability to determine copy number variation from SNP arrays has allowed for a new level of data analysis from existing studies [59]. With the detection of CNVs by SNP arrays combined with the benefits of SNP genotype data, hundreds of studies have performed large-scale CNV analyses. Additionally, it is thought that CNVs can be imputed using flanking or tagging SNPs, thereby eliminating the need for actual CNV detection. The basic premise is that if two samples differ in copy number, they are likely to either be tagged (i.e., SNP within CNV) or flanked by a SNP that will allow for the prediction of the CNV.

However, recent studies have demonstrated numerous limitations of SNP arrays for CNV detection and imputation. The ability of SNP arrays to tag or impute a CNV varies with many different factors. One factor is the number of probes present on an array and the resulting resolution for CNVs. SNP arrays have been constantly

expanding in their size from a few thousand probes to 4.3 million on the HumanOmni5-Quad BeadChip. The number of probes on SNP arrays provides lower resolution of CNVs than high-density CGH arrays. This lack in resolution partially arises from the limitation of only being able to place probes in locations that have a polymorphic variant within a population. Also, probes must pass a variety of stringent filters to ensure that they are unique in the genome ([24, 60, 61]). These filters often prevent probe placement in regions that are enriched for CNVs such as existing CNVs, segmentally duplicated regions, and repetitive elements.

Another limitation of SNP array genotyping for the detection of CNVs concerns the type of CNV to be imputed and its population frequency. For example, CNVs within SDs are often poorly tagged even when probes are located within the SD, possibly due to the repetitive nature of CNVs within the SD [24, 61]. Overall, CNVs located in SDs and tandem repeats with high mutation rates are unlikely to be imputed. Also, complex (i.e., recurring) CNVs are typically not imputed due to the formation of CNVs on multiple SNP alleles. Further analysis of CNVs has demonstrated differences within proportions of imputed CNVs depending on type of CNVs (i.e., duplications, deletions and complex CNVs). Deletions have the highest level of LD with SNPs because their locations are known (65%-81%) [24, 25, 60]. However, duplications and complex regions have much lower rates of imputation (24%), possibly because duplicated regions may be located non-tandemly with the original sequence. Therefore, SNPs used to impute duplications may be megabases ($\text{Mb}: 1 \text{ Mb} = 1 \times 10^6 \text{ bp}$) from the new copies of the DNA segments. Despite these limitations, approximately 61 to 80% of all CNVs are imputed or tagged by a SNP [24, 25, 60]. A final limitation of SNP imputation is the definition of LD between a SNP and a CNV. The limited number of studies investigating LD between

SNPs and CNVs show that very few CNVs have an r^2 greater than 0.8 (9%). However, approximately 50% of copy number polymorphisms (CNPs), which are polymorphic CNVs within a population (allele frequency >1%), have perfect LD with a SNP [24, 25, 60]. Since approximately 76% of CNVs are rare (i.e., <1% allele frequency), many CNVs will not be imputed. The inability to accurately impute CNVs with flanking or tagged SNPs leads to inaccurate genotyping of haplotypes. As a result, integrated CGH and SNP arrays are commonly being used [61]. Theoretically, this combined approach should increase the accuracy of genotyping, particularly in segmentally duplicated and repetitive regions. However, many of the limitations still exist, such as the probes placement and tiling density on an array. The inclusion of millions of SNVs from genome sequencing on a population level could further expand CNV and SNP analyses.

WHOLE-GENOME SEQUENCING

With numerous limitations in both CGH and SNP array analyses, the creation of massively parallel sequencing (next-generation sequencing), which is capable of generating billions of short sequence reads, may replace the use of arrays for genotyping. Whole-genome sequencing data can be quickly mapped to a reference genome and analyzed for SNVs, INDELs and CNVs. Massively parallel sequencing uses a common sequencing adaptor that is ligated to the ends of fragmented DNA [62, 63]. These adapters allow sequencing from one end (single-end) or both ends (paired-end) of the DNA fragments. These short reads (<150bp) can be sequenced on a variety of sequencing machines, such as the Illumina GAII and HiSeq. The GAII is older, but can achieve around 35 million reads (single-end) or 70 million (paired-end) from a single reaction (lane). The reactions occur on a flow cell that can run up to 8

independent reactions, allowing for the generation of billions of bp of sequence from a single run of the machine. The HiSeq improved the sequence generation by allowing for 150 million (single-end) or 300 million (paired-end) reads to be generated from a single lane.

The vast amounts of data collected from next-generation sequencing can be mapped to a reference genome using a variety of programs such as CLC Genomics, BWA, MrFAST, and Bowtie [64-66]. Once the reads are mapped, the genomes can be analyzed for variants (e.g., SNVs and INDELS) relative to the reference genome. Also, if paired-end sequencing is performed, programs can identify large INDELS, translocations, and inversions in a genome [67]. Many programs have been developed to identify CNVs based on independent and comparative analyses of sequence read-depth. The independent analyses perform depth corrections based on known biases such as GC content and masking repetitive regions [64, 68, 69]. The majority of programs perform comparative analyses of read-depths by selecting a genome as a reference and comparing the other to it. This method is more accurate because it accounts for other technical biases that may occur throughout the sequencing process.

While the use of next-generation sequencing is useful, it is not without limitations. First, a reference genome is needed to map sequencing data. Furthermore, because of the high level of repetitive elements and segmental duplications within a genome, it is almost impossible to map reads to the entire genome, even if a large number of reads are generated [70]. Second, there are the problems associated with identifying CNVs. Many CNVs occur in regions that are hard to map, such as repetitive regions and SDs. Also, many of the CNV programs lack the resolution to identify small CNVs with low false discovery rates (FDRs) [68, 69, 71, 72]. In general, the FDR of

structural variants from sequencing is much greater than CGH arrays. Last, most variant algorithms cannot detect novel sequences in a genome that may arise from MEIs and deletions in the reference genome.

The limitations of next-generation sequencing may be somewhat mitigated through the generation of *de novo* assemblies from next-generation sequencing data. However, this process is still expensive and requires a variety of libraries with varying insert sizes ranging from 0bp to > 40kb. Despite this, the *de novo* assembly of a genome often relies on other genomes to fully assemble the contigs into scaffolds [73, 74]. Regions of perfect or near perfect sequence duplications (segmental duplications and CNVs) add another level of complexity for *de novo* assembly. SDs, by definition, have greater than 90% sequence homology with another region in the genome, with some recently formed SDs having greater than 98% similarity [74]. In some cases, it is almost impossible to fully assemble these regions, leading to the collapse of many SDs into a single copy region. Although next-generation sequencing and *de novo* assemblies may alleviate many of the problems faced by array technologies, limitations of sequencing technologies prevent a comprehensive analysis of genome variation and structure.

EPIGENETIC GENE REGULATION

While the improvement of sequencing and other DNA techniques facilitates the discovery and investigation of genetic variation, these methods lack the ability to analyze non-genic differences that have been shown to play a major role in gene regulation. The regulation of transcription, silencing, and splicing can occur by a variety of post-translational modification to DNA and histone proteins. There are a variety of histone modifications (e.g., H3K4, H3K9, H3K27, H3K36, H3K79, and H3K20). Histones

can be modified by the presence of one to three methyl groups (monomethylation, dimethylation and trimethylation). Each modification is enriched within discrete genomic regions with a wide range of functional roles. Histone 3 lysine 4 trimethylation (H3K4me3) is one of the most commonly studied modifications due to its enrichment at transcriptional start sites (TSSs), promoters and, to some extent, intergenic regulatory elements (REs). The presence of H3K4me3 is often used to identify TSSs and is associated with rates of gene transcription. While the presence of H3K4me3 often indicates active gene transcription, it is also known to be present at inactive genes that may become active, such as genes that are silent in the G0 phase but active in G1 [75]. This modification binds TBP-associated factor 3 (TAF3) and recruits RNA polymerase II, thereby leading to active gene transcription. The presence of H3K4me3 can be cell specific and highly variable between individuals. Other histone modifications with both active and repressive roles are enriched for 5' regions of genes (e.g., H3K9me1, H3K20me1 and H3K79me1), across the gene with enrichment at 3' regions of genes (e.g., H3K36me3), expressed exons (H3K4me3, H3K36me3, H2BK5me1, H4K20me1, and H3K79me1), and introns (H3K4me1 and H3K36me1) [76-81].

In addition to histone methylation, DNA methylation plays a significant role in gene regulation. DNA methylation is a repressive modification that is typically associated with transcriptionally silenced genes [35, 82-87]. When DNA methylation is present at gene promoters, the transcriptional activity depends on the CpG content of the methylated promoter. In general, methylation of high CpG promoters leads to gene inactivation, while methylation of low CpG promoters can lead to activation or repression [88, 89]; however, DNA methylation can also silence transcriptional repressors, thereby resulting in enhanced gene transcription. While DNA methylation is

associated with repression, its presence within a gene (e.g., intron and exons) is indicative of active transcription. DNA methylation can occur over genes, exons, promoters, and large genomic regions [35, 82-87]. A commonly known example of DNA methylation is the silencing of genes on the inactive X chromosome in females [35, 82-87]. The role of DNA methylation on transcriptional regulation and splicing can also cause a variety of diseases [84-86, 90].

Recently, the Encyclopedia of DNA Elements (ENCODE) project released a large-scale analysis of epigenetic marks from 147 human cell types and 12 epigenetic modifications. Overall, the ENCODE consortium was able to identify 399,124 enhancer-like and 70,292 promoter-like regions [35]. The large number of regions resulted in 95% of the genome lying within 8 kb of protein-DNA interactions and 99% within 1.7 kb of a biochemical event. To determine the potential biological role of regulatory regions, 4,492 SNPs were compared with phenotypes in the National Human Genome Research Institute GWAS catalogue. Regulatory regions (e.g., transcription-factor-occupied and DNase-hypersensitive sites) were enriched for GWAS SNPs, with up to 31% and 71% of SNPs residing within or near the regions. Collectively, the combination of GWAS SNPs with regulatory information allows for the fine mapping of casual quantitative traits and diseases [35, 91-93].

CNVs AND TRAIT ASSOCIATION

The vast amount of human data demonstrates that both inherited and novel CNVs can be causal for disease and traits. In humans, CNVs have been estimated to cause at least 18% of the differences in genes expression for both normal and pathological samples [94]. CNVs are also suggested to be a major contributor to the missing heritability of complex traits in both humans and domestic animals.

While there have been many large scale association studies in domestic animals, very few have resulted in the identification of causal mutations for quantitative traits. The usage of GWAS, CGH, SNP genotyping, and next-generation sequencing has resulted in the identification of many disease- and trait-causing variants in domestic animals. The majority of these studies have evaluated dogs where more than 30 structural variants have been linked to diseases. Traits such as skin wrinkling (16.1kb complex duplication) and hair ridges (133 kb duplication) in dogs have been linked to CNVs [7, 95]. Recent studies are just beginning to identify CNVs linked to traits in other species, such as horses (i.e., coat color and early graying) [96, 97]. The limited number of known causal variants nevertheless demonstrates that CNVs contribute to phenotypes in domestic animals and provide the foundation for future studies in cattle populations.

BOVINE EVOLUTION

There are approximately 1.3 billion cattle worldwide, with many of the animals serving as a major source of beef and milk for millions of people [98]. The modern cattle populations existing today are the product of two independent domestication events and thousands of years of both natural and artificial selection [99]. These events resulted in the creation of multiple species, subspecies, and breeds with a vast array of distinctly unique traits.

The *Bos taurus taurus* and *Bos taurus indicus* subspecies of cattle, which represent the majority of cattle in existence today, diverged over 250 thousand years ago and were then independently domesticated on different continents approximately 10 to 12 thousand years ago [100-102]. *Bos taurus* cattle are the primary beef and dairy breeds in North America and Europe, whereas *Bos indicus* cattle are the primary

breeds in South America, Africa, and Central and South Asia. Different effective population sizes and selective pressures between *Bos taurus* and *Bos indicus* cattle resulted in distinct morphological and metabolic traits among cattle of these subspecies.

Although the traits of cattle vary by breed, *Bos indicus* cattle typically are characterized by a humped back, long face, steep upright horns, large dewlaps, unique set of coat colors (white/grey), and often a particular type of coat (slick coat) [103, 104]. *Bos taurus* cattle have their own unique set of morphological characteristics, such as diverse coat colors and patterns, absent horns in some breeds, enhanced milk yield, and superior carcass quality. Thermotolerance, parasite and pathogen susceptibility, blood pressure, and the onset of sexual maturity also differ between *Bos taurus* and *Bos indicus* cattle [103, 105-108]. These and other trait difference between the subspecies may influence susceptibility to disease. Infectious diseases, such as bovine respiratory disease and mastitis, result in large financial losses for the beef and dairy industries [109, 110]. Therefore, identifying variants underlying differential susceptibility to infectious diseases between *Bos indicus* and *Bos taurus* may facilitate genetic selection strategies for enhanced immunity and reveal novel immune pathways.

IMPORTANCE OF UNDERSTANDING TRAITS IN CATTLE

The bovine industry in the United States has been rapidly growing in terms of retail value even though the numbers of cattle and meat consumption have constantly decreased [111]. The increase in demand and loss in supply has placed an emphasis on lowering costs while improving production traits in cattle. The traditional focus on improving traits led to increases in milk and meat in cattle; however, losses due to diseases represent major costs to the agricultural industry. In Europe, the financial losses from diseases in cattle cost the cattle industry annually an average of \$82 per

cow [110]. In the United States, cattle and calf death resulting from illnesses totaled \$2,352,899,000 in 2010 [109]. The inclusion of costs associated with prevention and treatment of diseases in cattle would drastically increase the economic losses. While these costs to the agricultural industry are economically significant, the impact on human health may be even more significant. General usage of antibiotics, have led to agricultural industries collectively being the leading user of antibiotics. The overuse of antibiotics has been linked to the rapid increase in antibiotic resistance in pathogenic bacteria in humans [112]. The increase in prevalence of antibacterial resistant pathogens will continue to make it harder to treat diseases in animals, including people. Eventually, disease-causing agents could become resistant to all or a majority of antibiotics, making treatment difficult if not ineffective, such as occurs with extensively drug-resistant tuberculosis (XDR-TB). Therefore, the ability to improve the immune traits in cattle is not only necessary economically and for the welfare of cattle, but also for human health.

Table 1.1 Trends in the beef cattle industry

Year	# Cattle (million)	Retail Value (billion)	Beef Consumption (billion pounds)
2006	96.3	\$71	28.1
2007	96.6	\$74	28.1
2008	96.0	\$76	27.3
2009	94.5	\$73	26.8
2010	93.9	\$74	26.4
2011	92.7	\$79	25.6

BOVINE GENOMICS

The release of the completed genome assembly for a single *Bos taurus* cow (Hereford) in 2009 provided the first detailed look into the genomic structure of cattle [100]. The bovine reference assembly, Baylor version 4.0 (bosTau4.0), was created from contigs with an N50 (i.e., 50% of contigs are at least this length) of 48.7 kb and was estimated to represent 92% of the actual genomic sequence. Analysis of the genomic sequence demonstrated its complex structure, which consists of SDs, evolutionary breakpoints, repetitive elements, and transposed repeats. The bovine genome contains at least 124 evolutionary breakpoint regions (EBRs), of which 100 were specific to ruminants. The EBRs are enriched for LINE-L1 and LINE-RTE elements. In addition to EBRs, 3.1% of the genome is affected by 1,020 high-confidence SDs, which are enriched at EBRs. Furthermore, 778 of the SDs overlap genic regions. The high sequence identity of these duplications (98.7%) suggests they were recently formed [100]. Even though the genes affected by SDs are enriched for gene families involved in reproduction, many other genes are also affected. While the release of the bovine genome provides a treasure trove of data, it has expanded the abilities of genetic and genomic studies in cattle.

The search for the genetic basis of traits in cattle has spanned decades and yielded at least 71 Mendelian traits with underlying causative variants; however, most mutations are linked to diseases (<http://omia.angis.org.au/>). The major traits of interest, such as immunity and production, are quantitative traits that are unlikely to be caused by a single gene or mutation. Early studies of immunity and production traits have identified quantitative trait loci (QTL) that are associated with specific traits. Since the first QTL study to identify associations with milk production in 1995, hundreds of studies

have investigated more than 407 traits [113]. According to the Cattle QTL database, there are currently 5,920 publicly accessible QTLs for 407 traits. While these studies provide much insight into regions of the genome controlling complex traits in cattle, most studies have been unable to identify a single causative mutation due to the lack of resolution. Furthermore, the practice of sequencing exons within QTLs does not account for CNVs or variation at non-coding regulatory elements [114].

Recently, several studies have used array-based methods to examine CNVs in various cattle breeds. The first reports of CNVs in cattle utilized a whole-genome 385K CGH array designed by Nimblegen, which has been used in at least four studies [115-118]. Despite the low level of resolution (CNV > 24 kb), over 200 CNVs have been identified using the Nimblegen array, with enrichment of CNVs occurring primarily in EBRs and SDs. Collectively, these identified CNVRs account for approximately 1% of the genome and most of them (70%) overlap the coding portion of the genome. A second-generation, high-density array was recently generated by Nimblegen, consisting of 6.3 million probes tiled across the genome [119]. The HD Nimblegen array has a resolution of one probe per 420 bp, the highest resolution of any commercially available array fabricated to date. Fadista *et al.* used this HD array to investigate CNVs in 20 cattle and identified 204 CNVRs ranged in sizes from 1.7 kb to 2 Mb, which were enriched at SDs (20%) [119]. While this study was able to improve on the resolution of CNVs in cattle, CNVs below 1.7 kb in length still remained undetectable.

SNP arrays have also been used to identify CNVs in cattle. The bovine 50K SNP Beadchip array has been used to identify CNVs in six studies [120-125]. However, the large spacing (50 kb) between probes prevents the identification of CNVs smaller than 50 kb in length. Due to the poor resolution, the SNP array identified an average of

two CNVs per sample. Large scale studies using the 50K SNP array yield hundreds of CNVs throughout the bovine genome, with many occurring within SDs and QTLs. The majority of CNVs identified by the SNP array are imputed by a flanking SNP on the array [120]. As with CGH analyses, there are several programs used to detect CNVs from SNP arrays (e.g., PennCNV and CNVPartition). Comparison of these programs shows congruent results with 94% of the CNVs overlapping between the algorithms.

The BovineHD SNP array which contains over 770,000 probes has been used to perform two large scale CNV studies in cattle [126, 127]. Collectively 770 cattle have been investigated to identify CNVs as small as 1,018 bp in length [126]. Despite the small size of a few CNVs, the average and median sizes were much larger, 42 kb and 16 kb, respectively. Also, it was shown that the false discovery rate (FDR) for these CNVs is as high as 23.5% [127]. Therefore, while the BovineHD array provides an improvement from the 50K SNP arrays and low density CGH arrays, high FDR and low resolution still prevent the identification of small CNVs below 1 kb in length.

USE OF NEXT-GENERATION SEQUENCING TO EXAMINE GENETIC VARIATION IN CATTLE

In addition to array technologies, next-generation sequencing has been used to identify genetic variants in cattle [71, 128-132]. These studies have provided a glimpse into genetic variants at a genome level with low coverage analysis of SNVs, INDELs, and CNVs. Analysis of a Fleckvieh bull by next-generation sequencing identified over 2.4 million SNPs and 115,371 INDELs; however CNV analysis was not performed [128]. Additional genomic sequences of Angus, Holstein, Kuchinoshima-Ushi, and Nellore cattle (4X – 22X) have also been performed to identify SNVs, INDELs, and CNVs [71, 129-132].

There are a number of programs for identifying SNVs, INDELs, and CNVs from next-generation sequencing data; however discrepancies exist among the variant detecting algorithms [64, 67-69, 72, 133-135]. Zhan *et al.* used four different pipelines to identify variants in a single next-generation sequenced genome and found that only 48% of the SNVs were consistent among the pipelines [132]. A comparison of CNV algorithms has not been performed to date, but several programs exist including, CNV-Seq, MrFAST, Control-FREEC, and Breakdancer [64, 67-69, 72]. Comparisons among sequencing based CNV detection algorithms, SNP, and CGH arrays show little overlap among the methods (23%) [132]. Collectively, the use of next-generation sequencing to discover genetic variants has revealed high levels of genetic variation among cattle. However, there are no published studies that have compared genetic variation between breeds of *Bos taurus* and *Bos indicus* cattle.

THE NEED FOR A CENTRALIZED VARIANT DATABASE

Currently, there is no centralized database of genomic variants in cattle, such as the Database for Genomic Variants, the Human Gene Mutation Database and the 1000 Genomes project. These resources are essential for the investigation of candidate causal mutations. For example, variants from large studies can be filtered to remove any that have been previously shown to exist in control individuals because it is unlikely that these variants are disease-causing. Furthermore, variants shared between animals with different phenotypes likely do not underlie that particular phenotype. A centralized database would be a place to have all the collective genetic variants for an organism, and, if accompanied by phenotypic data, the database could be used to screen candidate variants for a myriad of Mendelian and complex traits.

DISSERTATION STRATEGY FOR CHARACTERIZING CNVS IN THE BOVINE GENOME

Collectively, the present study represents a comprehensive genetic analysis of the bovine genome using our custom analysis pipeline integrating whole-genome sequencing, high resolution CGH, CNV imputation by SNP arrays, and epigenetic profiling with previously known data (Figure 1.1). Utilizing our pipeline, this study demonstrates that CNVs are a major source of genetic variation in cattle populations. These variants and novel regulatory elements, along with previously identified variants, will become an essential part of future GWAS and linkage studies. In addition, this study provides insight into the limitations of CNV imputation by commercially available SNP arrays. Understanding these limitations will allow for more comprehensive GWAS and linkage studies with the ability to interpret CNVs commonly missed during traditional haplotype imputation. Overall, this study demonstrates that CNVs and SNVs within regulatory elements may underlie many phenotypic differences between *Bos taurus* and *Bos indicus* cattle.

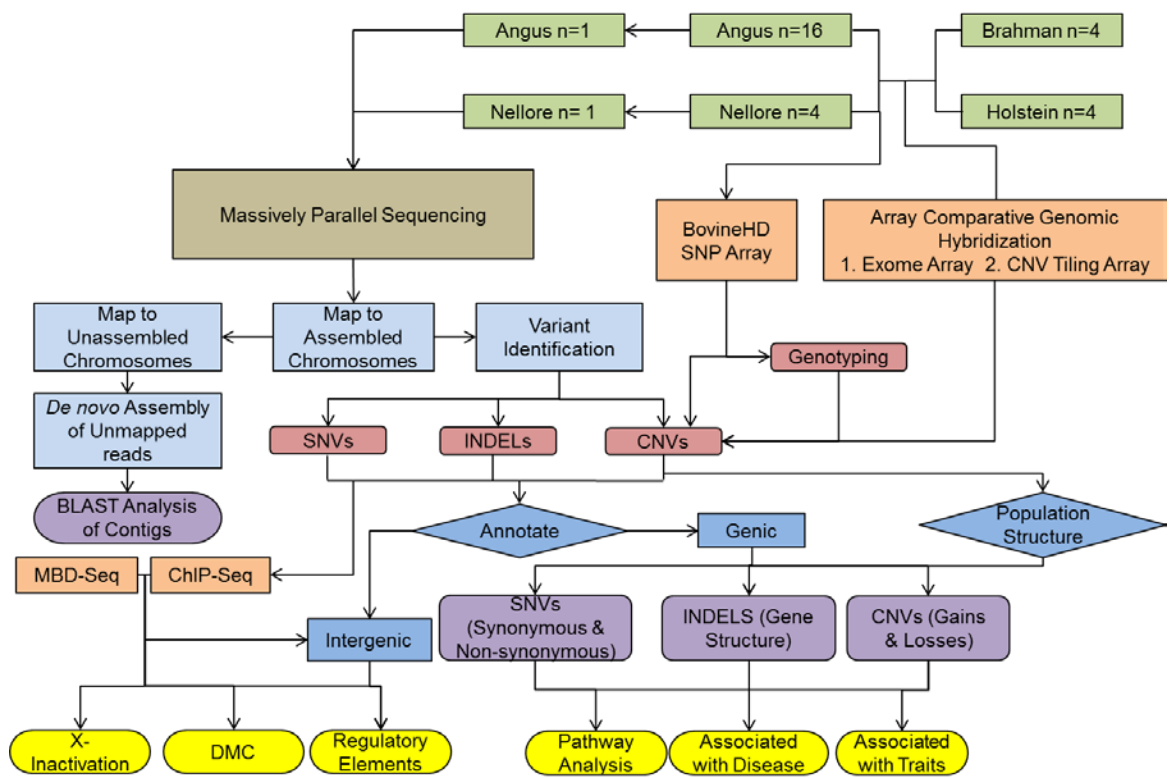


Figure 1.1 Genomic and epigenomic analysis pipeline

CHAPTER II

CGH ANALYSIS OF CNVS

INTRODUCTION

The two subspecies of cattle, *Bos taurus taurus* (*Bos taurus*) and *Bos taurus indicus* (*Bos indicus*), differ in a large number of metabolic and external traits. The breeds within these subspecies are known to be highly diverse in production and reproduction, leading to numerous genomic studies aimed at better understanding and improving these traits. Recently, the discovery that copy number variants (CNVs) underlie numerous phenotypes in other species has led to a focus of improving the understanding CNVs in cattle. To date, however, the extent to which CNVs exist in the bovine genome remains largely unknown.

The *Bos taurus* cattle have been the primary focus of CNV studies due to their dominance in both the dairy and beef industries in the United States. The previous 12 studies have used a combination of different CGH designs, SNP arrays, and whole-genome sequencing [71, 115, 116, 118-121, 123, 125, 126, 131, 132]. While, in theory, these methods should capture the majority of variants in a genome, they have a variety of limitations with unknown implications. Nearly all of the known 2,579 CNV regions (CNVRs) are from *Bos taurus* cattle with only one study investigating a single *Bos indicus* sample [71].

The first bovine CNV study utilized a 385K genome tiling array, designed by Nimblegen, to identify CNVs in three Holstein bulls [117]. Due to the poor resolution of this array design, only large CNVs could be investigated. This study was followed by a larger study, also using the 385K CGH array, that identified 229 CNVRs (52 located on ChrUn.) within 90 cattle, including 8 *Bos indicus* and 82 *Bos taurus* [116]. Despite array

resolution limitations, these studies demonstrate that CNVs are located through the bovine genome, with enrichment in SDs. Soon after the first reports of CNVs in the bovine genome, new studies have expanded the number of known CNVs and improved the CNV resolution through the use of 2.1M and 6.3M probe Nimblegen tiling arrays. These studies have been combined with lower resolution studies using large numbers of cattle in conjunction with the Bovine 50K SNP array. Most recently, a study utilized the BovineHD SNP Beadchip to investigate CNVs in cattle. In total over 2,457 cattle, predominately *Bos taurus*, have been studied for large CNVs [71, 115, 116, 118-121, 123, 125, 126, 131, 132, 136, 137].

The focus on *Bos taurus* cattle has led to only two studies attempting to associate CNV's with immunity. The first study provided an initial look into the role of CNVs on parasite resistance by looking at five related Angus samples [115]. The use of a 385,000 probe genome tiling CGH array allowed for the determination of large CNVs greater than 24 kb in length. The authors conclude that the 20 CNVs, many of which overlap immune-related genes, could play a role in parasite resistance and susceptibility in Angus cattle. In a follow-up study using the bovine 50K SNP array on 472 Angus samples, autosomal 2,724 CNVs cluster into susceptible and resistant groups [137]. By overlapping the affected genes with those known to be differentially expressed due to *C. oncophora* infections, several candidate gene CNVs including *ABO*, *IGLL1*, *LRRC17*, *MAMDC4*, *OAS1*, and *SERPINA5*. CNVs also affect other immune related genes including *WC1.1*, *LSP1*, *ABBC4*, and *TXNTD2*; however the selection of these genes is not significant ($p > 0.05$). Collectively, these studies demonstrate that CNV genes are enriched for immune-related processes, but various

limitations prevent the identification of CNVs causal or associated with parasite resistance in the Angus cattle.

The utilization of SNP arrays for CNV studies, especially the 50K Bovine SNP Beadchip, provides an extremely limited understanding of CNVs due to the placement of probes in unique regions containing polymorphic SNPs and the low tiling density. The 50K SNP array has an average probe spacing of nearly 50 kb, thus missing any small variants. As an alternate method of CNV detection, CGH arrays are highly accurate, while their resolutions have varied [138]. The first bovine array CGH design, tiling the entire genome with 385,000 oligos, is unable to detect small CNVs [117]. While later designs utilize two million and six million probes, they still lack the ability to detect small genic mutations due to poor exon tiling. Also, many studies use the same array designs, perpetuating the limitations into future studies. Finally, the design and selection of the probes on commercial arrays lead to another set of issues. These designs use extremely stringent criteria for probe selection, limiting probe selection within segmentally duplicated regions, tandem repeats, and regions of high GC content. These filtering criteria result in a loss of coverage across many important functional regions such as exonic tandem repeats and gene families.

We have minimized many of the limitations of previous array based CNV studies through the identification of CNVs in four breeds of cattle, belonging to both subspecies, with custom CGH designs. Our custom high-resolution exome oligonucleotide array provides the highest resolution analysis of coding CNVs to date. This analysis resulted in the identification of 754 CNVRs as small as 223 bp in length that were located in regions such as exons, exonic tandem repeats, and gene families. Copy number variants were significantly enriched in several biological classes including immunity and

defense. Additionally, we demonstrated a high level of shared variant genes across multiple species, suggesting predisposition to CNVs and potential functional roles.

METHODS

DNA Sample

We collected whole blood from 28 cattle belonging to four breeds: Angus, Holstein, Brahman, and Nellore. The white blood cells (WBC) were collected from the whole blood using two digestions in a red blood cell lysis solution containing 0.5 mM ethylenediaminetetraacetic acid (EDTA) and water. DNA was extracted from the WBCs using a standard phenol chloroform method consisting of two washes with phenol-chloroform-isoamyl (PCI), one wash with chloroform, isopropanol, and a final precipitation with 70% ethanol. The samples were suspended in Qiagen EB buffer and stored at 4°C (Qiagen Sciences, Germantown, MD) (Appendix 2.1).

Comparative Genomic Hybridization Array Designs

In order to identify exonic copy number variants in the bovine genome, we developed a high-density, exon focused, tiling array consisting of 418,336 unique oligonucleotide probes (Appendix 2.2). The Ensembl56 Biomart web service was used to extract the exonic and un-translated region (UTR) sequences with 40 bp of sequence flanking both 5' and 3' ends (www.ensembl.org). We selected unique oligonucleotides (oligos) of 60 bp in length using Oligowiz2 with the following parameters: Aim length = 60 bp; Max oligo length = 60 bp; Minimum oligo length = 45; cross-hyb minimum homology = 75%; cross-hyb length = 15 bp; cross-hyb max homology = 98%; cross-hyb length = 80%; and Minimum distance between oligos = 25 bp [139, 140]. Following probe selection by Oligowiz2, we stringently filtered the probes placed on the array using the following filters: Cross-Hyb > 0.2; Melting Temperature (C°), 72.5 – 84;

Folding > 0.26; Complexity > 0.2. Our final 418,336 oligonucleotides were checked to ensure uniqueness by random comparison to the reference genome using BLAT [141]. The final array design consisted of 418,336 oligos across 5' UTRs, 3' UTRs and coding exons of 411 micro RNAs (miRNA), 126 miscellaneous RNAs (miscRNA), 264 ribosomal RNAs (rRNA), 549 small nuclear RNA (snRNA), 460 small nucleolar RNAs (snoRNA), and 18,129 protein coding genes. On average, each gene was represented by approximately 20 oligonucleotides with one probe every 93 bp. The final array design was submitted to Agilent's eArray webserver for printing via Agilent's 60-mer SurePrint Technology (Agilent Technologies Inc., Santa Clara, Ca.).

In order to better define breakpoints of CNVs and identify additional variants flanking genes, we designed a CNV tiling array. All CNVs identified using the exome array in the Angus and Nellore cattle (Angus1-3, Nellore1-4) were merged into CNVRs. An additional 500 kb of flanking sequence was added to all regions. Chromosomal coordinates were uploaded into the eArray web service for probe selection against the bosTau4.0 assembly. The final probe set consisted of 414,700 unique oligos across 598 Mb of sequence. The final design was printed on Agilent's 2 x 400k array format using Agilent's 60-mer SurePrint Technology (Agilent Technologies Inc., Santa Clara, Ca.).

Array Comparative Genomic Hybridization (aCGH) Methods

We performed CGH to identify CNVs against a single reference Angus genome (Angus-4) (Appendix 2.3). DNA was sheared using a Sonic Dismembrator 500 and purified with an Invitrogen Purelink PCR Kit (Invitrogen, Carlsbad, CA). The sheared genomic DNA from the reference was labeled with Alexa Fluor 555 and all other samples were labeled with Alexa Fluor 647 fluorescent dyes using the BioPrime Plus Labeling module (Invitrogen, Carlsbad, CA). We mixed the reference and experimental

DNA and denatured with 25 μ l Cot-1 DNA (Invitrogen), 26 μ l Agilent 10X Blocking Buffer and 130 μ l 2X High-RPM hybridization buffer prior to hybridization at 65°C for 20 hours. Array slides were washed in Agilent wash buffer-1, wash buffer-2 and finally in acetonitrile. We scanned the slides at a 2- μ m resolution with an extended dynamic range (XDR) of 0.05 using an Agilent High Resolution Microarray Scanner 62505C (Agilent Technologies Inc., Santa Clara, CA). Agilent's Feature Extraction 10.7 software was used to extract data from the scanned images and perform quality control checks (Agilent Technologies Inc.). We imported the data into Agilent's Genomics Workbench v5 and used the Aberration Detection Module 2 (ADM-2) with a threshold of 6, bin of 10 and a centralization threshold of 6 to identify CNVs in respect to the reference Angus sample. CNVs were required to have an average \log_2 ratio of 0.5 across at least three consecutive probes. The \log_2 ratios were used to group CNVs into three classes: less than a 2:1 ratio, greater than or equal to a 2:1 ratio and homozygous deletions. Homozygous deletions were identified using \log_2 ratios of at least 2.5, at least three consecutive probes, and signal intensities equal to the background signal. Those CNVs in a heterozygous state (2:1) were identified as having log ratios ranging from 0.7 to 2.5, based on the average log ratio across the X chromosomes of male vs female comparisons. The maximum p value for all call groups was 10^{-10} . CNVs that met the previous criteria and at least 223 bp in length were considered high confidence. Any CNVs meeting the calling criteria with lengths less than 223 bp were considered low confidence (See CNV confirmation section for full explanation).

The CNV focused tiling array was used to confirm the breakpoints of large CNVs and identify additional intragenic variants flanking genes. The CNV tiling array was used to further investigate two Angus (Angus-2 and Angus-3), one Holstein (Holstein-2), two

Nellore (Nellore-2 and Nellore-3), and one male Brahman (Brahman-1) sample by comparison to the same reference used with the exon arrays. The preparation, hybridization and analysis procedures from the exon arrays were used for the CNV tiling arrays.

CNV Confirmation

Small insertion/deletion variants within genes were confirmed through the design of standard PCR primers flanking the regions using Primer3Plus and UCSC In-Silico PCR tools ([142], www.genome.ucsc.edu). Standard PCR protocols were used to confirm the smallest CNV that can accurately be detected by the exome array. The PCR products that possessed and lacked the CNV were inserted into a pCR2.1-TOPO plasmid and grown in Top10 chemically competent cells following manufacturer's TOPO TA Cloning protocol (Invitrogen, Carlsbad, CA). The selected white colonies were grown overnight in Lysogeny broth (LB) with kanamycin. Plasmids were isolated from the cultures using a QIAprep Spin Miniprep kit (Qiagen Sciences, Germantown, MD). The inserts were confirmed using both standard PCR with the CNV primers and *EcoR1* digestions. Plasmids containing and lacking the CNV were sequenced by the Texas A&M University DNA Technologies Core Laboratory using Sanger sequencing. The sequences were aligned to the bosTau4.0 genome using the UCSC Genome Browser, BLAT, and ClustalW (<http://www.ebi.ac.uk/Tools/msa/clustalw2/>).

The accuracy of the array designs was determined through several steps of confirmation involving 20 affected genes and 16 intergenic regions. Using Primer3 Plus and UCSC In-Silico PCR tools, we designed primers for standard PCR and quantitative PCR (qPCR) of genomic DNA (Appendix 2.4). Nine of the genic and sixteen intergenic

primers were used to confirm homozygous deletions using standard PCR procedures (**Appendix 2.5**).

We performed qPCR, using SYBR GreenER qPCR Supermix, to confirm differences in copy number and deletions by comparison to a reference gene (*GAPDH*) (Invitrogen). Samples possessing and lacking the CNVs were selected for qPCR analysis to determine relative copy number ratios and confirm calls by the exon arrays. All samples were analyzed simultaneously in triplicate to ensure an equivalent comparison of copy number. All qPCR reactions were performed by creating a 10- μ l mix with 25 nanograms (ng) DNA, SYBR GreenER, water, and primers (Invitrogen and Sigma-Aldrich). The samples were analyzed using an ABI 7900HT Fast Real-Time PCR machine to determine the cycle threshold (CT) values. The $\Delta\Delta$ CT method was used to determine the relative copy number changes by comparison to *GAPDH* [143].

Genomic Characterization of CNV Content

The level of copy number variation within chromosomes and the genome was determined from the total bases affected by CNVRs within each group (breeds, subspecies, and all samples). We determined enrichment values per chromosome by dividing the length of CNVRs by the total length of entire assembled chromosomes, all genes, all exons, and regions covered by the array. We chose to use the enrichment of genes covered by the array since intergenic regions were not tiled. Then, the total length of all CNVRs was divided by the length of the genes tiled to determine the genome enrichment. Enriched chromosomes were identified if their percent enrichment was greater than the percent enrichment of the entire genome.

CNVs were compared to known elements including segmental duplications (SDs), quantitative trait loci (QTLs), CpG islands, tandem repeats, conserved regions

(phastConserved elements), and known CNVs. All CNVs were overlaid with segmentally duplicated regions (WSSD, WGAC, and high confidence SDs) in the bovine genome using ANNOVAR [144, 145]. The CNVs were compared to QTLs from the Cattle and Bovine QTL databases [146, 147], http://genomes.ersa.edu.au:8080/bovineqtl_v2/). All CpG islands, tandem repeats, and conserved regions were downloaded from the UCSC genome browser. Known CNVs were compiled from all published cattle CNV studies and overlapping regions were merged to create a database of known CNVRs in cattle.

Analysis of CNV Content in the Bovine Umd3.1 Assembly

The bosTau4.0 chromosomal locations of all probes on the exome CGH array were converted to Umd3.1 positions using the liftOver tool from the UCSC genome browser (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>). The conversion caused 2,358 probes to be either unmapped or placed on un-assembled contigs. The remaining 415,978 probes were successfully placed on assembled chromosomes. A new exome array map file for the Umd3.1 genome was created in eArray. Using the new map file, all array images were re-extracted in Feature Extraction in order to create new data files with Umd3.1 coordinates.

The remapped data were analyzed in Genomics Workbench using the procedure from the bosTau4.0 analysis. Genomic content of CNVs in the Umd3.1 assembly was characterized as in bosTau4.0 analysis. Additionally, assembly related differences in CNV content were determined through the comparison between the bosTau4.0 and Umd3.1 analyses.

Functional Analysis of CNVs

We annotated the CNVs against the Ensembl, RefSeq and Human orthologous genes using the ANNOVAR software program [145]. The genic and exonic content within each CNV was ascertained from the annotation. The biological functions of genes affected by CNVs were characterized using the DAVID Functional Annotation Tool with the default settings [148, 149]. The resulting biological process terms were further grouped by similarities in function to identify enriched biological processes. Statistical significance (p value) of enriched groups was determined with the Fisher's combined probability test from the DAVID Functional Annotation Tool's p value (Equation 2.1).

$$X^2 = -2 \sum_{i=1}^k \log_e(p_i)$$

k=number of p values being combined

Degrees of freedom = 2k

$p_{i=p}$ = p value of sample 'i' to be combined

Equation 2.1 Fishers' combined probability test

All genes affected by CNVs were converted to RefSeq gene symbols and compared to the Online Mendelian Inheritance in Animals (OMIA) and Online Mendelian Inheritance in Man (OMIM) databases (<http://www.ncbi.nlm.nih.gov/omia> ; <http://www.ncbi.nlm.nih.gov/omim>). The genes with known phenotypic associations were identified for future studies into traits influenced by CNVs.

Population Analysis of CNVs

We performed cluster analysis of all samples using the Genesis clustering software [150]. The \log_2 signals for all probes within CNVs for each sample were imported into the Genesis software for clustering using the following parameters: Pearson correlation, hierarchical clustering, and complete linkage (Graz University of Technology: Institute for Genomics and Bioinformatics). The level of CNVs shared among the samples was determined through the comparison of CNVs from each sample. Sharing of CNVs was classified by the number samples with each CNV; for example, a CNV was classified as being unique if it was present in a single sample. Next, we compared the sharing of CNV genes among the samples and breeds by overlapping the gene lists from each sample and determining the total number of samples with CNVs in each gene.

We then performed separate global fixation index-statistic (FST) analyses of all CNV RefSeq and ensembl genes using the GenePop software (<http://genepop.curtin.edu.au>). The genotypes for CNV genes were predicted in all samples as follows: no CNV, Wt/Wt; $0.5 \leq \log_2 \leq 2.5$, Wt/Dup; $\log_2 \geq 2.5$, Dup/Dup; $-0.5 \geq \log_2 \geq -2.5$, Wt/Del; $\log_2 \leq -2.5$, Del/Del. The inter-population differentiation and global FST values were calculated for breeds and subspecies using the following parameters: Option 6; allele identity (F-statistics) for all populations (global FST) and for all population pairs (inter-population differentiation FST); fit to Ln(distance); convert F-statistics to F/(1-F)-statistics; minimum distance between samples, 0.0001; number of permutations for Mantel test, 1000; and, diploid. Fixation Index-Statistic values were compared to genes known to be selected for in *Bos taurus* and *Bos indicus* cattle [102, 151].

Since it has been shown that many genes and regions may be predisposed to copy number changes, we further investigated those genes in cattle. A custom database of known CNVs and affected genes in cattle, horses, dogs, mice, and humans was created compared to our bovine CNV genes ([71, 115, 116, 118-120, 123, 131, 132, 137, 152-161], Doan *et. al.* unpublished, and <http://projects.tcag.ca/variation/>). All genes from each species were converted to human ensembl gene IDs for comparison among the species. The genes were overlapped to determine the extent of sharing between each group using Microsoft Excel and the Venny online tool (<http://bioinfogp.cnb.csic.es/tools/venny/index.html>). Functional groups with genes commonly affected by CNVs (i.e., present in three or more species) were identified using the DAVID Functional Annotation Tool.

RESULTS

CGH Array Design

Given the preponderance of CNVs identified between the Angus and Nellore cows and the potential of these variants to have large effects on phenotypes, we examined CNVs occurring in cattle representing two *Bos taurus* (Angus and Holstein) and two *Bos indicus* (Nellore and Brahman) breeds. A combination of high-density exome and CNV tiling CGH arrays were used to identify and characterize CNVs in the cattle.

A custom exome focused CGH array was designed using ensembl annotated exons and UTRs in the bosTau4.0 reference assembly. The 418,336 unique probes densely tiled more than 95% of protein- and 75% of RNA-coding genes in the ensembl56 annotation database (411 micro RNAs [miRNA], 126 miscellaneous RNAs [miscRNA], 264 ribosomal RNAs [rRNA], 549 small nuclear RNA [snRNA], 460 small

nucleolar RNAs [snoRNA], and 18,129 protein coding genes (Figure 2.1, Table 2.1). The focus on tiling one probe every 93 bp across the exome resulted in an average of 41bp between each probe. On average, each gene was represented by approximately 20 oligonucleotides, providing the highest resolution of the bovine exome via aCGH. Furthermore, all probes were annotated against ensembl genes, CpG islands and tandem repeats and compared against the commonly used bovine 385K and 6.3M CGH designs by Nimblegen. These data clearly confirmed the advantage of our design in the identification of exonic CNVs, including those within in potential regulatory CpG islands and tandem repeats (Table 2.2).

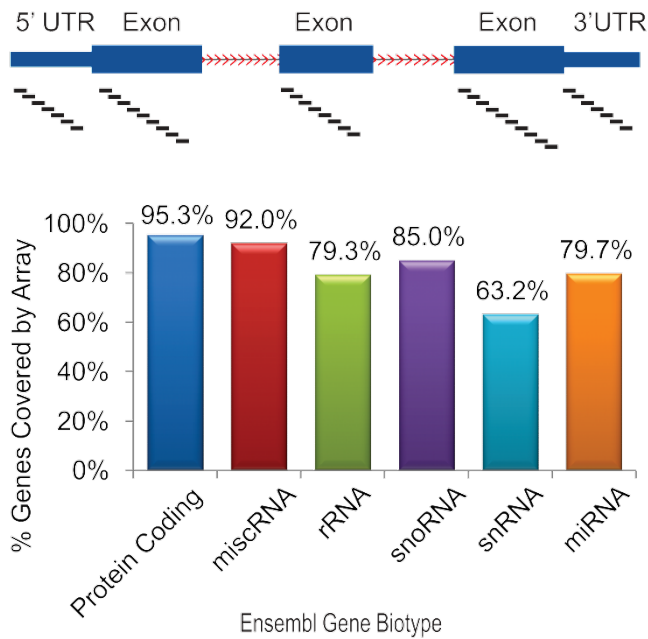


Figure 2.1 Gene coverage of bovine exome CGH array

Table 2.1 Statistics of exome CGH array design

Total Probes	Average Probe Length	Max Probe Overlap	Average Melting Temperature	Regions Covered	Exon Resolution	EnsGene Coverage	Protein Coding Genes	RNA Genes
418,336	52 bp	25 bp	79.6° C	Exons	93 bp	19,939	18,129	1,810

Table 2.2 Comparison of exome CGH and Nimblegen tiling arrays

Region	Total Probes			Probes Overlapping Islands			Probes Overlapping CpG			Probes Overlapping Tandem Repeats		
	Nimblegen 6.3M	Nimblegen 385K	Exome 400K	Nimblegen 6.3M	Nimblegen 385K	Exome 400K	Nimblegen 6.3M	Nimblegen 385K	Exome 400K	Nimblegen 6.3M	Nimblegen 385K	Exome 400K
Exon	122,273	7,809	306,517	15,404	944	19,975	1,545	65	3,100			
5' UTR	3,711	3,263	8,402	2,050	550	2,995	56	44	63			
3' UTR	19,848	3,968	103,168	556	85	1,506	197	36	726			
Intron	1,719,794	103,980	77	19,227	969	2	24,307	1,348	0			
Intergenic	4,529,468	259,957	172	32,377	1,229	0	59,842	3,683	3			
Total	6,395,094	378,977	418,336	69,614	3,777	24,478	85,947	5,176	3,892			

Exome CGH Accuracy and Resolution

The identification of CNVs throughout the exomes of 28 cattle (Angus, n=16; Nellore, n=4; Holstein, n=4; and Brahman, n=4) was initially performed using the custom exome CGH array (Figure 2.2). The array data were filtered to remove probes with saturated and abnormally high signal intensities. These regions consisted of highly repetitive elements such as pseudogenes, retrotransposons, LINES, SINES, and tandem repeats. For example, a CNV identified in the 3' UTR of the ensembl annotated *GPR137B* gene was entirely composed of a SINE (ART2A). Due to the errors associated with identifying and mapping highly repetitive probes to a specific location, probes were removed using a signal cutoff at three standard deviations above the average signal (100,000). The ADM-2 algorithm was tested to determine the best settings by increasing: the minimum number of continuous probes required in a CNV (3 to 10), the minimum \log_2 ratio for a CNV (0.25-0.5), and bin sizes (6, 8, and 10). These tests revealed that increased filtering and detection settings resulted in a much greater loss of CNVs than did loosening the stringency.

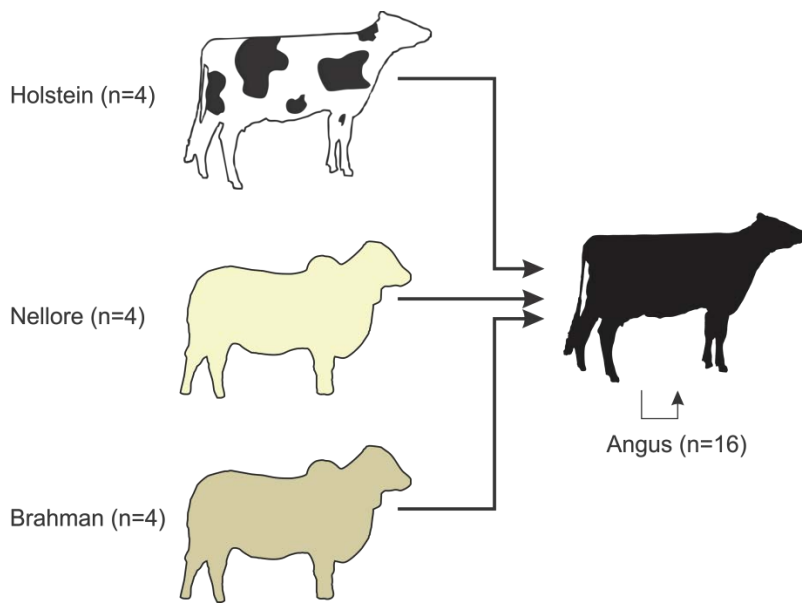


Figure 2.2 Samples used on exome CGH analysis

The accuracy and resolution of our custom array design, using the selected settings (3 consecutive probes, bin size of 6, and minimum \log_2 ratio of 0.5), was further tested using several methods. First, a self-self hybridization of the reference Angus cow predicted the false discovery rate (FDR) ranged from 0% to 6.2% (Table 2.3). The average \log_2 ratio across the X chromosomes of male versus female hybridizations (0.75 averaged over 12,196 probes) was used to determine the value of a 2:1 ratio of gene content. Homozygous deletions were characterized as CNVs with an average \log_2 ratio ± 2.5 [162]. The minimum size of CNVs confidently identified was determined by PCR and Sanger sequencing of six small variants ranging in lengths from 98 bp to 364bp. Of the variants, those with lengths of 98, 122, 223, and 364 bp were confirmed, while CNVs with lengths of 100 and 105 bp were unable to be confirmed (Figure 2.3-2.5). Despite two of four variants around 100 bp in length being

correctly identified, we chose to increase the minimum length for high confidence CNVs to 223 bp. Smaller variants were retained and analyzed separately. The accuracy of CNV calling was further tested by qPCR confirmation 100% of (20 of 20) CNV genes, including nine affected by homozygous deletions.

Table 2.3 False discovery rates of CNVs by classification

Length Criteria	# Probes	Log ₂ Range	p value	# of Calls	FDR
100bp - 222 bp	3-4 Probes	$0.5 \leq \log \text{ ratio} \leq 0.7$	10^{-10}	11	6.2%
	> 4 Probes	$0.5 \leq \log \text{ ratio} \leq 0.7$	10^{-10}	53	
	3-4 Probes	$0.7 \leq \log \text{ ratio} \leq 2.5$	10^{-10}	196	
	> 4 Probes	$0.7 \leq \log \text{ ratio} \leq 2.5$	10^{-10}	66	
	3-4 Probes	$2.5 \geq \log \text{ ratio}$	10^{-10}	34	
	> 4 Probes	$2.5 \geq \log \text{ ratio}$	10^{-10}	3	
≥ 223 bp	3-4 Probes	$0.5 \leq \log \text{ ratio} \leq 0.7$	10^{-10}	16	3.2%
	5 Probes	$0.5 \leq \log \text{ ratio} \leq 0.7$	10^{-10}	1,268	
	3-4 Probes	$0.7 \leq \log \text{ ratio} \leq 2.5$	10^{-10}	325	
	5 Probes	$0.7 \leq \log \text{ ratio} \leq 2.5$	10^{-10}	1,123	
	3-4 Probes	$2.5 \geq \log \text{ ratio}$	10^{-10}	54	0%
	5 Probes	$2.5 \geq \log \text{ ratio}$	10^{-10}	134	

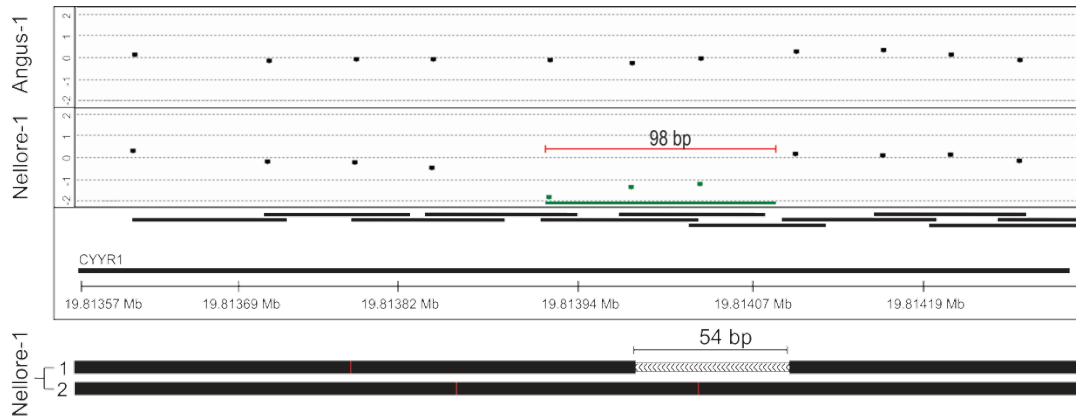


Figure 2.3 Identification and Sanger sequencing confirmation of a 98bp loss within 3' UTR of CYYR1 in Nellore-1

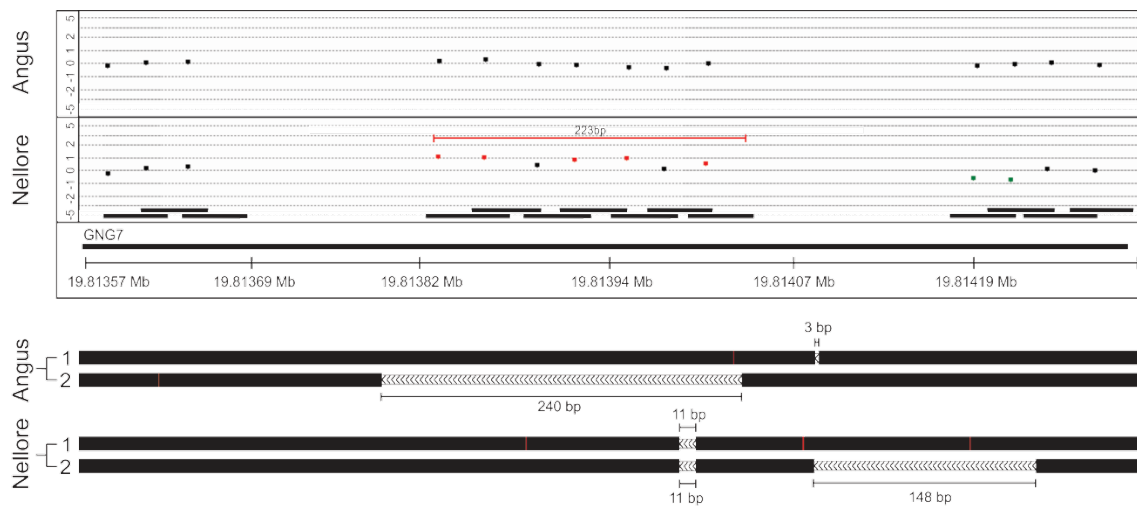


Figure 2.4 Confirmation of CNV used as smallest size for CGH analysis

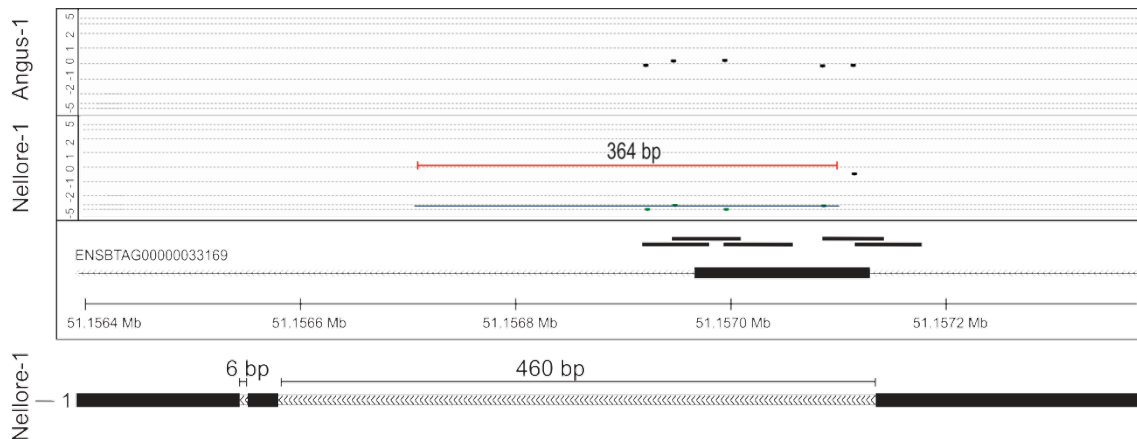


Figure 2.5 Confirmation of 264bp homozygous deletion affecting the majority of a single exon.

The ability of the exome array to identify large and complex CNVs was tested through the analysis of two large tandem CNVs on chromosome five. The variants were identified as being two regions totaling nearly 500,000 bp (Figure 2.6A). Combined, the CNVs affected 13 known and predicted protein coding genes belonging predominately to olfactory receptors. The deletion of the first region (chr5:63,170,402-63,504,768) was found in 17 samples (15 Angus, 2 Holstein), while the second region (chr5:63,802,864-63,979,026) was found in nine (8 *Bos indicus*, 1 Angus) samples. Additionally, based on the log2 ratios, Angus-2 and Nellore-1 were heterozygous for the CNVs.

The analysis of the complex CNVR on chromosome five began by determining the approximate breakpoints of the CNVs. Therefore, through the standard PCR using 18 intergenic and four genic primer sets, we confirmed the homozygous deletions and better defined the breakpoints (Figure 2.6B). Additionally, Angus-2 and Nellore-1 were confirmed to be heterozygous for the deletion and produced a product by PCR.

Quantitative PCR confirmed the presence of both homozygous and heterozygous deletions (Figure 2.6C).

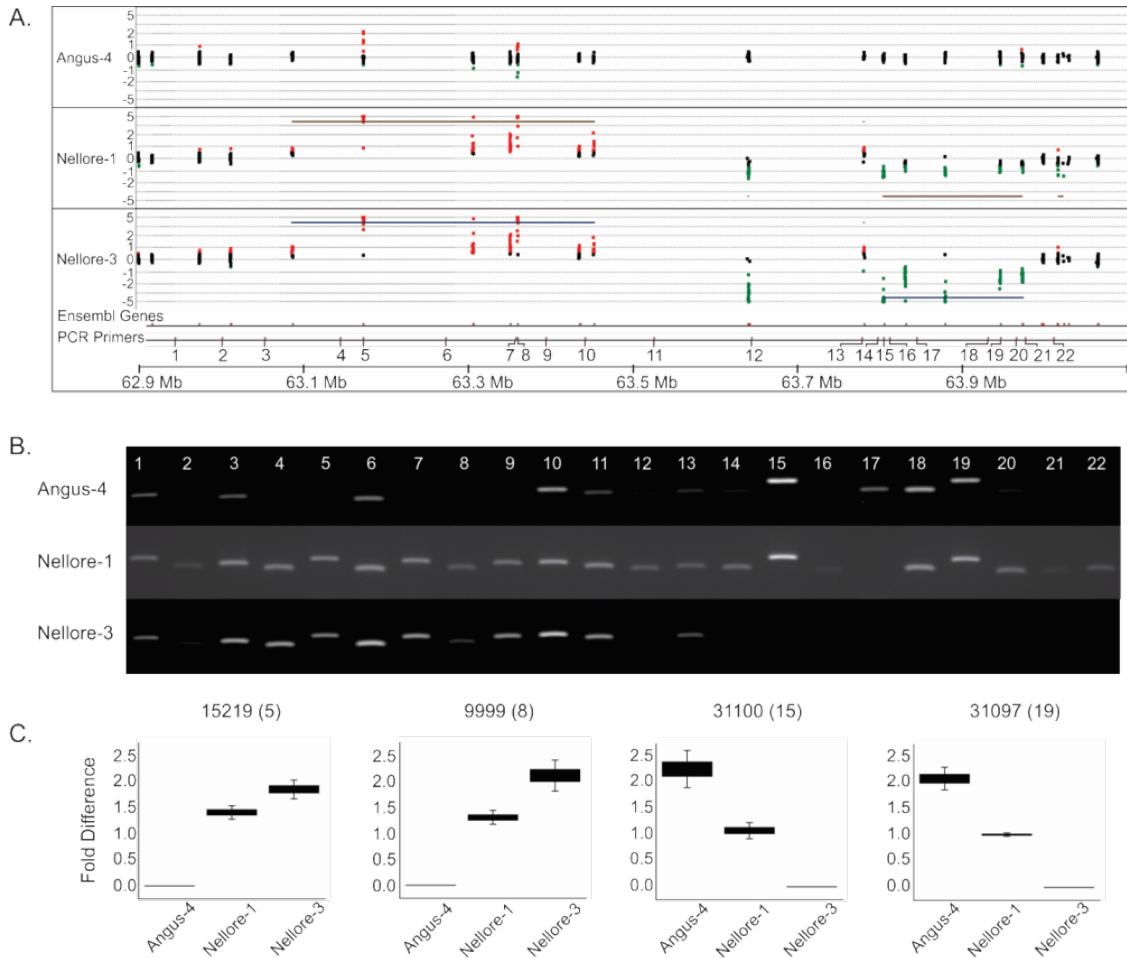


Figure 2.6 (A.) Identification and confirmation through (B.) PCR and (C.) qPCR of a large, complex CNV region on chromosome 5 of Angus and Nellore cattle

Genome Distribution of CNVs

Analysis of the 28 cattle samples revealed CNVs across all chromosomes with greatest significant ($p < 0.05$) enrichments on chr27, chr29, and chr7. Overall, it was estimated that CNVs affected approximately 4.9% of the bovine genome. Several types

of chromosomal regions were found to contain significantly increased enrichments of CNVs such as the MHC ($p=9 \times 10^{-52}$). We identified 754 (2,920 CNVs) high- and 162 (363 CNVs) low-confidence CNVRs. A slight increase in deletion regions was observed over duplications and an additional 58 complex CNVRs were identified (Table 2.4). While many of the small CNVs were likely real, we excluded them from the remainder of the analysis to decrease the FDR (Appendix 2.6). CNVRs were found to vary widely in length from 223 bp to 3,873 kb. The mean size of CNVs was 62.5 kb; however, CNVs were enriched for lengths of 632 bp and 32 kb (Figure 2.7). Additionally, 66% (497) of the CNVRs identified were novel. Finally, the comparison of average CNV densities by breed revealed that Nellore cattle possessed approximately a threefold increase in CNV content over the Angus cattle, when compared to an Angus reference.

Table 2.4 Identification of CNVs and affected genes across the bovine exome

Sample	CNVs (Genes)	2:1 CNVs (Genes)	Deletion CNVs (Genes)	CNVRs (Gain:Loss:Complex)	Common CNVRs
Angus (15)	1,157 (708)	565 (364)	86 (24)	375 (154:198:23)	5
Holstein (4)	267 (332)	112 (125)	15 (9)	179 (85:92:2)	9
Brahman (4)	370 (413)	182 (221)	24 (14)	226 (120:100:6)	11
Nellore (4)	1,126 (809)	589 (403)	63 (18)	515 (233:276:6)	95
<i>Bos taurus</i> (19)	1,424 (808)	677 (394)	101 (26)	414 (168:214:32)	2
<i>Bos indicus</i> (8)	1,496 (972)	771 (522)	87 (27)	585 (265:300:20)	8
Total (27)	2,920 (1,352)	1,448 (752)	188 (37)	754 (318:378:58)	2

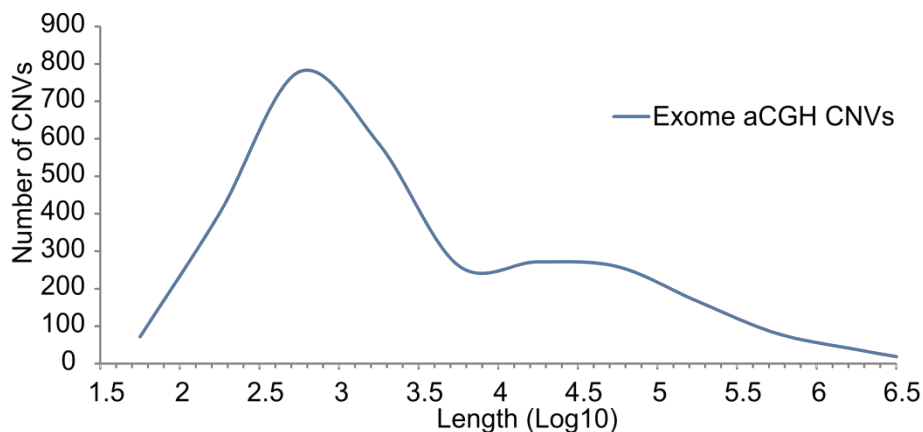


Figure 2.7 Enrichment plot of exome CNVs based on the log10 of variant lengths.

To better understand the genetic elements underlying copy number variation, CNVRs were compared to several regions known to affect genome structure and phenotypes. First, the comparison of SDs identified by WSSD and WGAC with CNVRs revealed that 39.4% overlapped at least one SD, while 50% of homozygous deletions were located within SDs. In addition, we found larger CNVRs were positively correlated with their overlap of SDs (Figure 2.8). While 44% of CNVRs contained tandem repeats, 22 were predicted to be a direct result of deletion/duplication events within a single exon. Interestingly, 79% of the CNVRs contained conserved regions, while only 26.5% contained CpG islands. Finally, nearly all CNVRs (96%) overlapped known cattle QTLs.

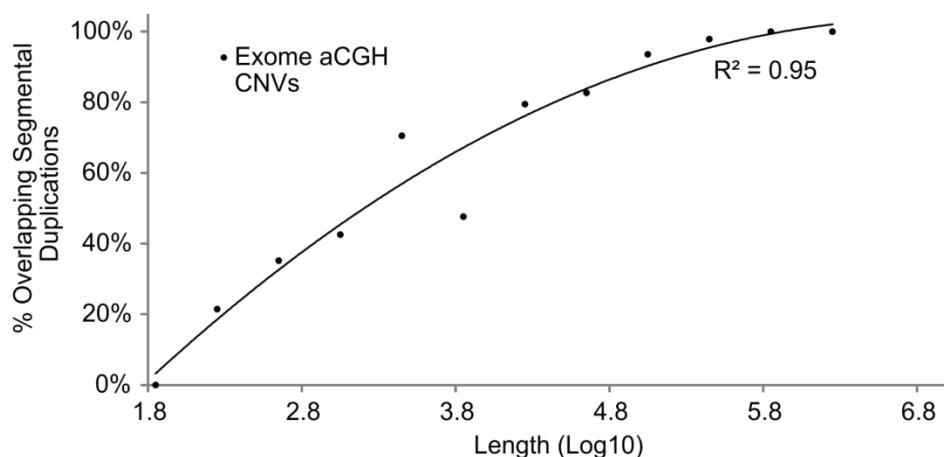


Figure 2.8 Plot of percentage exome CNVs overlapping segmental duplications

CNVs affected 1,352 protein- and RNA-coding genes from all biotypes, while the effects on the genes varied from a portion of a single element (UTR or exon) to entire gene clusters (Appendix 2.7). Of the single element CNVs, coding exons (418), 5' UTR (49) and 3' UTR (426) were affected. Additionally, tandem repeats were predicted to cause single element copy number differences in 22 genes. Homozygous deletions affected 37 genes, with single element deletions in 3' UTRs and coding exons of three and seven genes, respectively.

We investigated six animals (Angus, n=2 cows; Nellore, n=2 cows; Holstein, n=1 cow; and Brahman, n=1 steer) using the second-generation CNV tiling array. The array design covered nearly 600 Mb of sequence and included 483 CNVRs from the exome array (Table 2.5). The final array design covered 7,134 ensembl genes; however, unlike the exome array, the majority of the probes were located within intronic and intergenic regions and very few were located in CpG islands or tandem repeats (Table 2.6). The lower resolution of the array design allowed for the inclusion of more sequence regions.

Therefore we were able to investigate CNVs missed by the exome array (e.g., intronic CNVs and CNVs flanking genes).

Table 2.5 Probe statistics for CNV tiling array

Total Probes	Average Probe Length	Regions Covered	Resolution	EnsGene Coverage	Protein Coding Genes	RNA Genes
414,700	56 bp	+/-500kb CNVRs	1,443 bp	7,134	7,090	44

Table 2.6 Annotation of CNV tiling array probes

	Total Probes	Probes Overlapping CpG Islands	Probes Overlapping Tandem Repeats
Exon	15,950	1,173	109
5' UTR	6,471	745	46
3' UTR	8,939	163	61
Intron	154,985	1,340	1,179
Intergenic	228,355	1,223	1,871
Total	414,700	4,644	3,266

Using the focused CNV tiling array, we identified 411 CNVs (210 CNVRs), including 43 homozygous deletions. Given the resolution and lack of exon coverage of the CNV tiling array, many small CNVs were missed. However, the comparison of breakpoints of large CNVs between the exome array and the CNV tiling array indicated many were accurately predicted by the exome array with few large differences due to probe placement in the array designs. Of the 299 protein- and RNA-coding genes identified as CNVs with the tiling array, 80% were identified by the exome array. The

remaining 61 genes represented CNVs in regions not covered by the exome array (e.g., introns) or were removed due to stringent filtering. In total, 23 regions were located completely within introns. We identified 87 intergenic regions flanking genes that could not be identified using the exome array (Figure 2.9).

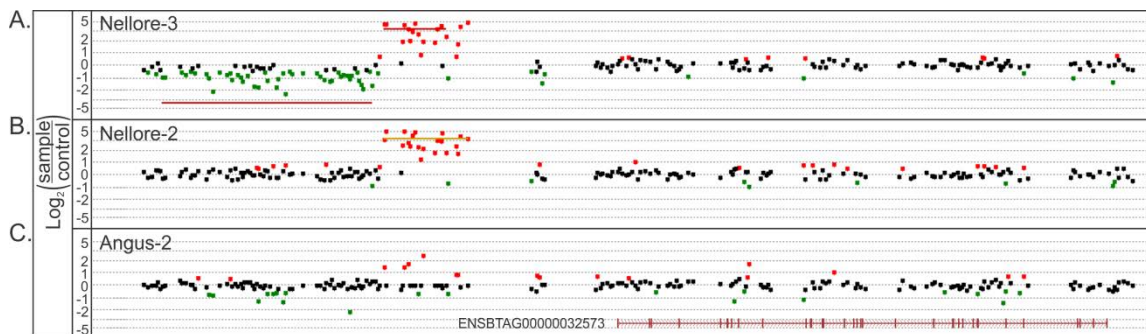


Figure 2.9 Identification of a complex CNV flanking the 5' region of a gene consisting of: (A.) flanking loss and gain, (B.) gains, and (C.) no CNV

Umd3.1 Analysis

The conversion of all bosTau4.0 (Baylor) exome probes to Umd3.1 (Maryland) coordinates resulted in the loss of 2,358 probes, but still allowed for the analysis of CNVs throughout the exonic portion of the assembly. In total, 725 CNVRs (2,742 CNVs) affecting 1,220 protein- and RNA-coding genes were identified in the Umd3.1 assembly. The comparison of these variants with those from bosTau4.0 revealed the majority of CNVs were present in both analyses, which resulted in the threefold increase of CNVs in the Nellore cattle over the Angus cattle. Overall, 376 genes did not overlap those from the Baylor assembly analysis. However, after accounting for CNVs filtered based

on our stringent criteria and genes not on assembled autosomes in the bosTau4.0 assembly, only 23 genes were unique to the Umd3.1 assembly.

Further comparison of CNVs and affected genes between bosTau4.0 and Umd3.1 assemblies revealed structural differences that prevented the identification of all CNVs in either assembly. For example, the CNV of the *CATHL1* and *CATHL4* antimicrobial genes was very different between the two assemblies (Figure 2.10). The CNV in the Baylor assembly consisted of *CATHL1* and *CATHL4*; however, the Umd3.1 assembly lacked the *CATHL4* gene. Within the cathelicidin region, the Umd3.1 assembly was nearly 20 kb shorter and lacked a predicted cathelicidin gene and *CATHL4*. Therefore, analyses on the Maryland assembly cannot identify a *CATHL4* CNV, which possibly prevented its association to immune traits. Furthermore, analyses on the Baylor assembly also missed genes affected by CNVs. For example, one major difference in the Umd3.1 assembly was the length of the X chromosome, where it was approximately 60.3 Mb larger than in bosTau4.0. Further investigation found that much of the unassembled Baylor sequence was actually located on the X chromosome of the Maryland assembly. While we were unable to ascertain copy numbers for these genes, some were actually identified within Umd3.1 CNVs because of densely tiled flanking genes. Despite the ability to identify some of these genes, it appeared that a comprehensive analysis of exonic CNVs would require assembly specific array designs.

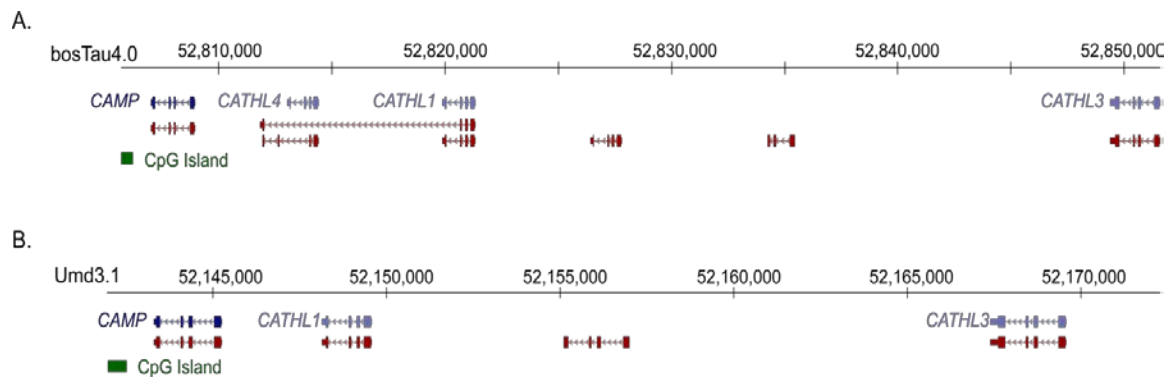


Figure 2.10 Comparison of the cathelicidin locus in the (A.) bosTau4.0 and (B.) Umd3.1 genome assemblies.

Functional Analysis

We found protein-coding genes, small nucleolar (snoRNA) genes, microRNA (miRNA) genes, small nuclear (snRNA) genes, miscellaneous RNA (miscRNA) genes, and ribosomal RNA (rRNA) genes were among the 1,352 genes affected by CNVs (Figure 2.11A). Overall, the majority (94.67%) of the genes with CNVs were protein-coding (Figure 2.11B). Additionally, genes affected by CNVs and homozygous deletions were both primarily enriched in processes involving immunity and defense, signal transduction, and sensory perception (Figure 2.11D-D). All statistics for BP analyses are reported in Appendix 2.8. We also found that unique CNVs were primarily enriched in processes involved in sensory perception ($P=0.005$) (Figure 2.11E). In comparison, we found an even greater enrichment for CNV genes in immunity and defense processes in the exome Umd3.1 analysis (Appendix 2.9). Collectively, these data indicated that CNVs were present in all biotypes and that biological processes regulating immunity and defense and signal transduction were enriched for CNVs.

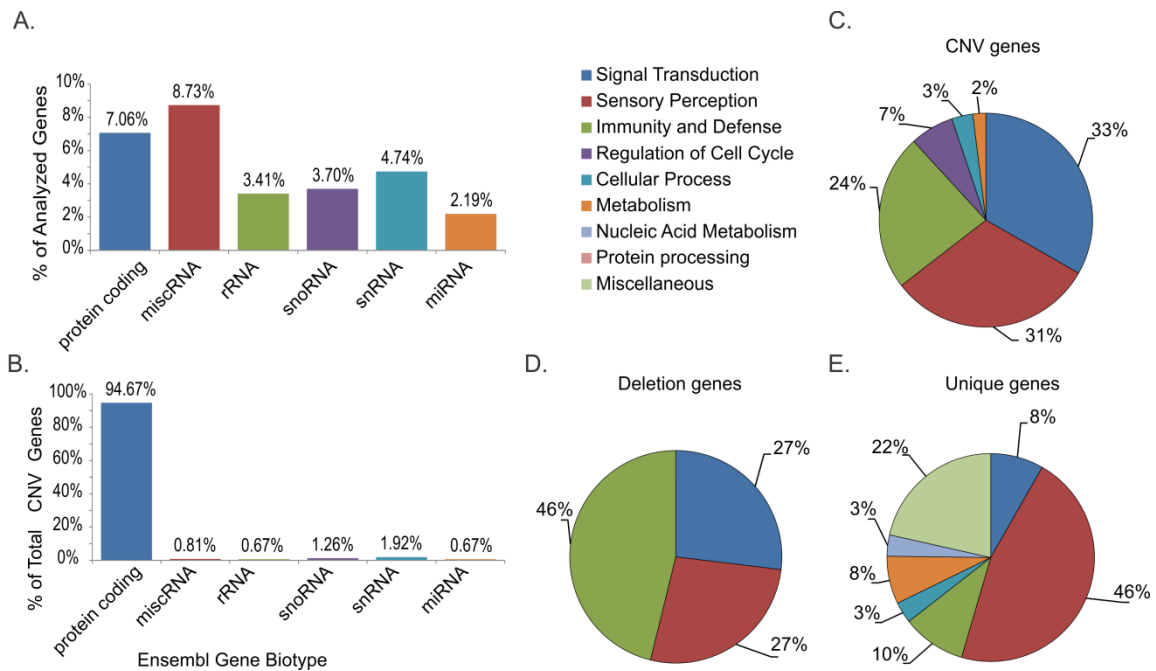


Figure 2.11 Functional analysis of genes affected by CNVs in the exome CGH array. (A.) Percentage of analyzed genes within each biotype affected by a CNV. (B) Distribution of biotypes within CNV genes. Biological process enrichment of: (C.) all CNV genes, (D.) homozygous deletion genes, and (E.) unique genes.

Next, we examined whether the CNVs were present at genes associated with Mendelian traits in animals and humans. Cross-reference of the genes affected by CNVs with the Online Mendelian Inheritance in Animals (OMIA) database returned 16 genes (Table 2.7). However, the comparison with the Online Mendelian Inheritance in Man (OMIM) database revealed that 135 genes were present in the morbidity map and 342 had OMIM terms (Appendix 2.10).

Table 2.7 CNV genes with associated OMIA terms

Gene	OMIA ID	Phenotype	Gene Description	Chr	Start	End	Location	Gain/Loss	Samples
AGL	1577	Glycogen storage disease IIIa	amy(1b)-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase	chr3	46,320,588	46,321,204	Exon	Gain	3
AGTPBP1	662	Motor neuron disease, lower	ATP/GTP binding protein 1	chr8	83,072,362	83,072,545	Exon	Gain	4
AF3B1	248	Neutropenia, cyclic	adaptor-related protein complex 3, beta 1 subunit	chr10	8,743,749	8,744,219	Exon	Loss	2
BMPR1B	383	Fecundity, Booroola	bone morphogenetic protein receptor, type IB	chr6	31,251,555	31,251,733	Exon	Loss	3
CFH	636	Membranoproliferative glomerulonephritis type II	CFH complement factor H	chr16	3,918,791	4,149,328	Genes	Loss	6
CNGB3	1365	Achromatopsia (cone degeneration, hemeralopia), AIMA1	cyclic nucleotide gated channel beta 3	chr14	73,435,528	74,973,151	Genes	Loss	1
CYP27B1	837	Vitamin D-deficiency rickets, type I	cytochrome P450, family 27, subfamily B, polypeptide 1	chr5	60,136,415	60,140,638	Genes	Loss	1
EDNRB	629	Megacolon	endothelin receptor type B	chr12	53,571,317	54,304,962	Genes	Loss	1
GALC	578	Krabbe disease	galactosylceramidase	chr10	101,100,924	102,594,090	Genes	Complex	2
GDF9	385	Fecundity, Iceland (Thoka)	growth differentiation factor 9	chr7	44,090,799	44,091,936	Exon	Gain	1
NPHP4	1455	Cone-rod dystrophy, Standard Wire-haired Dachshund	nephronophthisis 4	chr16	44,573,368	44,573,646	Exon	Loss	2
PSMB7	1454	Coat colour, harlequin	proteasome (prosome, macropain) subunit, beta type, 7	chr11	98,642,163	98,642,978	Exons	Loss	5
SERPINH1	1483	Osteogenesis imperfecta_Dachshund	serpin peptidase inhibitor, clade H (heat shock protein 47), member 1	chr15	54,315,009	54,315,248	Exon	Gain	1
SOD1	263	Degenerative myelopathy	superoxide dismutase 1, soluble	chr1	2,909,417	2,915,482	Genes	Loss	3
TEX14	1673	Spermatogenic arrest	testis expressed 14	chr19	8,794,337	8,794,692	Genes	Loss	2
UROS	1175	Porphyria, congenital erythropoietic	uroporphyrinogen III synthase	chr26	46,182,981	46,195,627	Exons	Gain	1

The cathelicidin-1 (*CATHL1*) and cathelicidin-4 (*CATHL4*) antimicrobial genes were found to be duplicated in all four Nellore cattle and three Angus cattle using the exome CGH array. The 16-kb duplication of *CATHL1* and *CATHL4* was confirmed by qPCR. These genes belong to a family of cathelicidin genes that originated from expansions from a single cathelicidin gene. The cathelicidin genes don't appear in OMIA; however, CAMP (the human orthologue of cathelicidin) was listed in OMIM. Cathelicidin is described to be an antimicrobial peptide that, upon maturation by cleavage between the signal and antimicrobial domains, is secreted by leukocytes and epithelial cells to promote inflammation, angiogenesis, wound healing, and tumor metastasis [163]. While more studies are needed, we suspect that the effects of gene duplications within this family would contribute to heightened innate immunity in cattle.

Additionally, growth differentiation factor 9 (*GDF9*) was affected by a duplication of exon 2 in a single Angus sample (Figure 2.12). *GDF9* is expressed in oocytes and is essential for ovarian folliculogenesis. Heterozygous mutations in this gene cause an increase in fertility while homozygous mutations cause sterility [164]. It is not clear whether the heterozygous duplication will have the same effect as a deletion on fecundity or if the sample has any distinct fecundity traits.

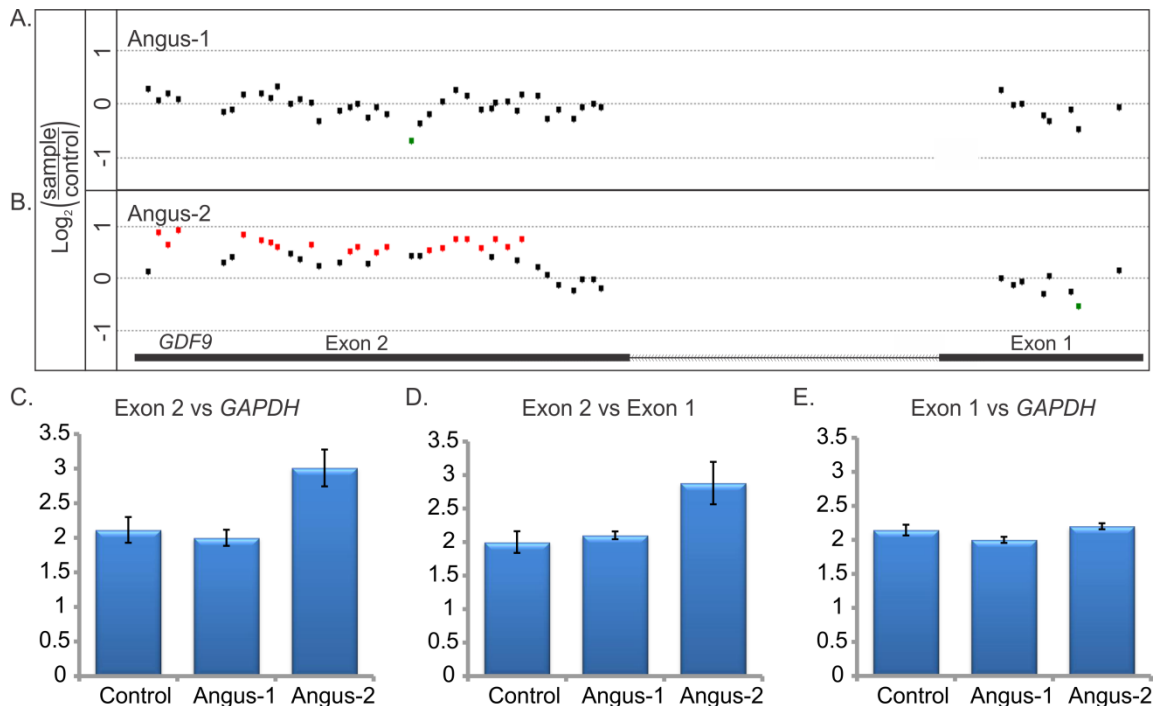


Figure 2.12 Exon resolution analysis of single exon CNV in *GDF9*. Identification of samples (A.) without and (B.) with exonic CNV. Confirmation using qPCR through the comparison of (C.) exon 2 with *GAPDH*, exon 2 with exon 1, and exon 1 with *GAPDH*.

The superoxide dismutase 1 gene (*SOD1*), encodes a protein that binds copper and zinc ions and is responsible for removing free superoxide radicals from the body. The *SOD1* gene contained a loss in the 3' UTR and exon-5 in all Nellore samples. Mutations in this gene are causal for degenerative myelopathy in dogs [165]. *SOD1* is also located within the polled locus in cattle and could be a candidate for polling in cattle [166-168].

We identified a small CNV in all Nellore samples located within the 3' UTR of the bone morphogenetic protein receptor, type IB (*BMPRI1B*). This gene (*BMPRI1B*) encodes a bone morphogenetic protein receptor involved in bone formation and

embryogenesis. Mutations in *BMPR1B* are associated with multiple ovulations and sterility in sheep [169]. Nellore cattle are slower to reach sexual maturity and have longer calving intervals; therefore, this CNV could play a role in fecundity traits of Nellore cattle [170]. However, these preliminary findings will require more investigation to elucidate the true effects on phenotypes.

Population Analysis

Hierarchical clustering of all probes within CNVRs indicated that CNVs were shared among animals and within breeds; however, further investigation of the variants revealed that 41% were unique to individual animals (Figure 2.13 A). Also, few variants were shared among all samples of a breed; Nellore contained the most shared regions (95) (Figure 2.13 B). Additionally, CNV lengths were positively correlated with the level of sharing observed between cattle (Figure 2.13 C). Regions less than 1 kb in length were more likely to be unique (60%) while CNVRs greater than 1 kb were more likely to be shared with another sample (70%).

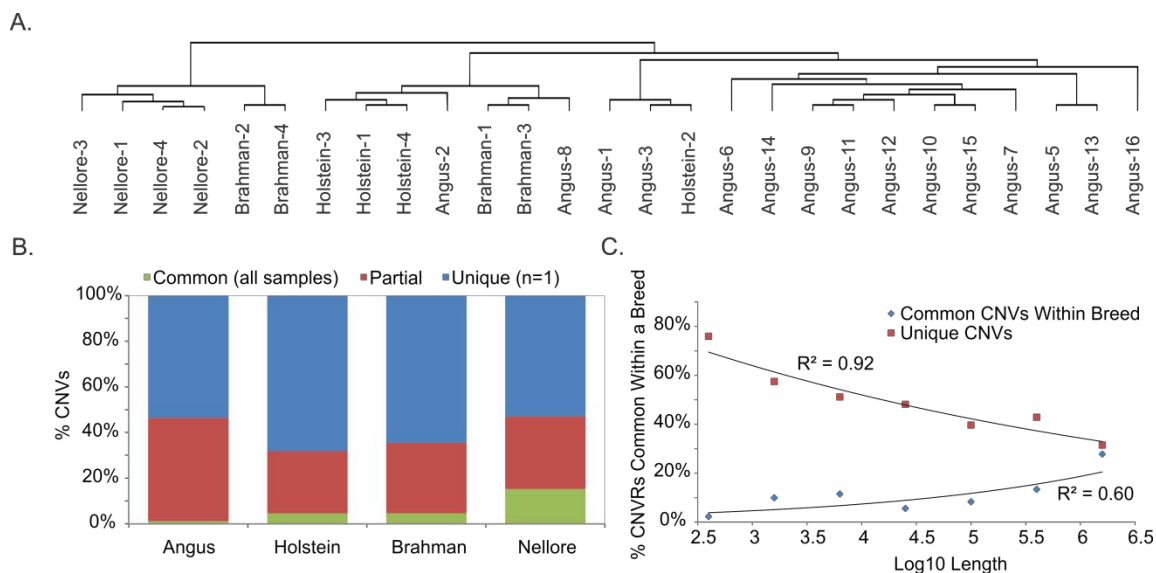


Figure 2.13 Population structures of CNVs among each breed. (A.) Hierarchical clustering of all probes within CNVs. (B.) Plot of overlap among CNVs of samples within each breed. (C.) Positive correlation of shared CNVs with increasing lengths.

As final step in the understanding the functional role of CNVs between breeds and subspecies, a global F_{ST} analysis of CNV genotypes identified genes (RefSeq and Ensembl) under selection. The global F_{ST} between the four breeds revealed numerous CNVs potentially under selection in different groups. The greatest F_{ST} values resulted from 21 genes with CNVs present only in the Nellore samples (e.g., *PSMB7*). *Bos taurus* samples possessed several genes with elevated F_{ST} values (e.g *FANCC* and *IGLL1*). Immune related genes (e.g., *YWHAZ*, *KLRF1*, *CATHL1*, and *CATHL4*) were under positive selection and existed at heightened frequencies in *Bos indicus* cattle (Appendix 2.11 - 2.12). Furthermore, several of the genes with high F_{ST} values were previously identified as regions of selection in cattle, including both *Bos taurus* and *Bos indicus* subspecies. However, many of the genes potentially under selection were

previously undetected, possibly due to previous limitations in identifying and genotyping CNVs.

Comparative Analysis

The phenotypic differences of recurring and unique CNVs could drive breed and individual specific traits in cattle. Therefore, in order to identify genes predisposed to copy number changes, we compared CNV genes in horses, dogs, cattle, and humans. Given the lack of CNV information for horses and the limited data available for dogs, we created custom CGH arrays for both species. The horse array was designed like the bovine array and focussed on the ensembl annotated exome [155]. In order to maximize the CNVs identified, we selected 15 divergent breeds and a donkey as an evolutionary outlier. Overall, 775 CNVRs ranging in sizes from 187 bp to 3.5 Mb affected 1,707 ensembl genes [155].

Our analysis of canine CNVs utilized a 400K array design focused on genes, including both exons and introns. While this design limited our resolution, we demonstrated the presence of intronic CNVs among the 252 CNVRs. Unlike the cattle and horse analyses, relatively few CNVs were found in each sample, suggesting much less structural differences in the breeds. Despite the lower level of CNVs, 437 genes were affected by CNVs.

All CNV genes we identified in horses, dogs, and cattle were combined with known CNV genes to create a database of CNVs and affected genes in humans, dogs, horses, mice, and cattle. The conversion of all genes into human ensembl orthologs resulted in the loss of highly divergent and species-specific genes. There were over 35,000 human genes affected by CNVs, but less than 4,000 for the other species, which suggested that many more shared genes have yet to be identified

(<http://dgvbeta.tcag.ca/dgv/app/home?ref=NCBI36/hg18>). Despite the possibility of numerous unknown CNVs and up to 95 million years of divergence between the species, 20 genes were found to be copy number variant in four species (Human, Dog, Cow, and Horse) and 4,335 are shared between cattle and at least one other species (Figure 2.14). Many of the genes (34%) that were shared with at least one other species were located within segmental duplications in the cattle genome. The remaining 537 genes possibly represented CNVs specific to cattle. The genes shared in all species were members of gene families such as olfactory clusters, immunoglobulins, and amylases. Genes shared between cattle and at least two species were enriched for biological processes involved in sensory perception (40%, $p=3.3 \times 10^{-12}$) and cellular processes (15%, $p=1.5 \times 10^{-6}$). The high level of CNV gene overlap between species and the apparent enrichment of specific classes of genes suggested a predisposition of genes to copy number changes.

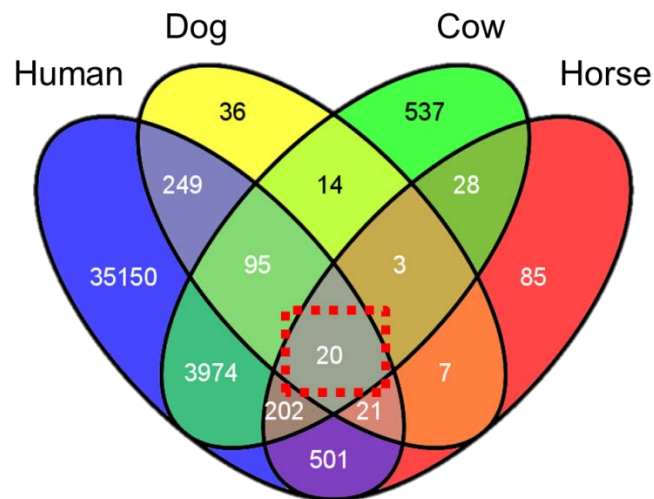


Figure 2.14 Sharing of genes affected by CNVs among diverse species

DISCUSSION

The bovine genomics field has spent billions of dollars attempting to identify causal mutations underlying economically important traits in dairy and beef cattle. While production traits have obvious benefits to agriculture and the economy, disease prevention and treatment costs the industry billions each year and is a contributor to antibiotic resistance. Unfortunately, in cattle, there has been a lack of success in identifying mutations including both single base and copy number variants. Therefore, the role of CNVs on phenotypes (e.g., immune function) remains largely unknown.

The previous attempts to identify CNVs in cattle through array CGH have largely focused on whole genome tiling methods. While these methods were good for identifying large genic and intergenic CNVs over 10 kb in length, they often lack the ability to identify variants below 1 kb and lack exonic coverage. While variation within intergenic regions may be functionally relevant, variants affecting the coding portions of genes may be more likely to affect proteins and phenotypes. Therefore we created the first exome focused tiling array specifically designed to cover the majority of coding and non-coding exons in the bovine genome with an average resolution of 93 bp. Our resolution and ability to identify variants arising from tandem repeats far exceeded previous studies. The benefits from our exome design allowed for the identification of many novel variants in the bovine genome.

Through the use of our custom exome CGH array, we characterized genetic variation within cattle breeds accounting for 4.9% of the genome. Unlike previous findings, we clearly described copy number changes enriched for sizes well below the previous resolutions. The positive correlation of CNV length with uniqueness suggests that smaller CNVs are occurring at much greater mutation rates than large CNVs.

Therefore, while many mutations may exist at predisposed sites such as tandem repeats and SDs, small novel CNVs may be influencing individual traits within breeds. Further study may elucidate the role of these variants on intra-population differences.

The role of CNVs on traits in cattle has yet to be determined however, with CNVs affecting regions as small as a portion of a single exon and tandem repeats within exons, the potential effects on phenotypes are drastically increased. Smaller CNVs have the potential to create novel proteins from in-frame mutations. Given the numerous methods in which a CNV can affect expression, the 1,352 genes affected by CNVs have the potential alter a wide range of phenotypes, including immune function. Further analysis of CNVs using FST, biological processes, and OMIA revealed many potential candidate genes such as the *CATHL1*, *CATHL4*, and *YWHAZ* genes for immune function. The extent of small exonic and whole gene CNVs suggests a vital role in phenotypic diversification both within and between breeds of cattle.

With costs of disease prevention and treatment rising, the understanding of the genetic differences influencing traits like immune function will be critical to the agricultural community as well as to the world's economy and food supply. While our study does not definitively identify causal variants for specific traits, we have clearly shown that CNVs of all sizes exist throughout the bovine genome. Until high resolution and focused studies characterizing genomic variation are completed, the utility of large level association and functional testing of candidate CNVs will be hindered. Therefore, we expect that this study will become a basis for future descriptive studies leading to large scale scans for specific traits. These technologies, in combination with whole-genome sequencing, will lead to rapid growth in bovine genomics.

CHAPTER III

WHOLE GENOME ANALYSIS OF *Bos taurus* AND *Bos indicus* COWS

INTRODUCTION

The divergence of *Bos taurus* and *Bos indicus* cattle more than 250,000 years ago led to adaptations for specific climates and agricultural functions [100-102]. The majority of cattle breeds used for milk and meat production in the United States belong to the *Bos taurus* subspecies, including the Angus and Holstein breeds. However, in regions with tropical climates, breeds belonging to the *Bos indicus* subspecies have become the predominate cattle. While all of these breeds belong to the same species, they are highly divergent in ancestry and phenotypes. Holstein cattle, the predominate dairy breed in the United States, are experiencing a rapidly increasing level of inbreeding due to a combination of artificial insemination and the desire to improve specific milk production traits [171-173]. Angus and other beef breeds were selected for meat quality and feed-efficiency traits. In contrast to human selective pressures for production, the Nellore cattle (*Bos indicus*) originated in India and were imported into Brazil. In Brazil, Nellore cattle were exposed to a large range of pathogens due to the tropical climate, resulting in a heightened immunity and an increased tolerance for extreme heat [105-107]. Overall, cattle breeds have been the target of a wide range of natural and human selective pressures, leading to a high level of phenotypic divergence.

The costs related to diseases account for major economic losses for the agricultural industry throughout the United States. For example, bovine respiratory disease (BRD) and Johne's disease are collectively estimated to cost the agricultural industry over one billion dollars annually from production losses, treatments and deaths

[174, 175]. Therefore, identifying the variants underlying disease susceptibility between *Bos indicus* and *Bos taurus* may allow for a better understanding of immunity and possibly better breeding strategies.

Despite over 400 genome-wide association studies (GWAS) in the past 20 years, the vast majority phenotypic differences in *Bos taurus* and *Bos indicus* cattle have no known causal mutations, but over 5,920 quantitative trait loci (QTLs) [146, 147, 176]. Prior to next-generation sequencing, many of these studies relied on SNP genotyping arrays and microsatellites. However, recent advancements in next-generation sequencing technologies allowed seven studies to sequence a total of ten *Bos taurus* (Angus, Holstein, Fleckvieh, and Kuchinoshima-Ushi) and two *Bos indicus* (Nellore) genomes [71, 128-132, 177]. The amounts of analyses in these projects have varied; however, the majority of studies focused on single animals for SNP detection. Additionally, one of the Nellore cows was used in the creation of a rough *de novo* assembly, not a reference mapping [178]. Recently, a study performed a comparative analysis of three Angus, one Holstein, and one Nellore genome using whole-genome sequencing; however, this study focused primarily on differences in copy number [71]. Despite the limited number of samples sequenced, it is clear that a large amount of genetic variation has yet to be identified.

Whole-genome sequencing allows for the identification of genic and intergenic variation that cannot be identified by CGH or SNP arrays [57, 74, 179]. The limited number and incomplete analyses of cattle genomes clearly demonstrates the need for a combined approach that includes the comparison of subspecies. Therefore, we performed whole-genome sequencing of a single Angus and Nellore cow to identify single base, small insertion/deletion, and copy number variants within each genome.

Through a functional and evolutionary analysis we were able to gain insights into the role of genetic variation underlying phenotypic differences between Angus and Nellore cattle.

METHODS

Whole-Genome Sequencing

DNA was isolated from ear notches of an Angus and Nellore cow using a standard phenol chloroform method consisting of two washes with phenol-chloroform-isoamyl (PCI), one wash with chloroform, one wash with isopropanol, and a final precipitation with 70% ethanol. The samples were suspended in Qiagen EB buffer (Qiagen Sciences, Germantown, MD). For the construction of sequencing libraries, we first sonicated high-quality genomic DNA by pulsing 3 times for 15 seconds per pulse at 14% power using a Sonic Dismembrator 500 (Fisher Scientific, Pittsburg, PA) and purified with an Invitrogen Purelink PCR Kit (Invitrogen, Carlsbad, CA) (Appendix 3.1). The DNA was blunt end-repaired, adenylated, and ligated with paired-end adaptors, according to the manufacturer's protocol (Illumina, San Diego CA). The prepared library was resolved on a 2% low range agarose gel and a 2-mm section of DNA with an insert size of 231 bp was extracted from the gel (Qiagen Sciences, Germantown, MD). The library was enriched according to the manufacturer's protocol (Illumina, San Diego CA). The size and concentration of the sequencing library was determined by PCR, polyacrylamide gel electrophoresis (PAGE), and through the use of the Agilent 2100 Bioanalyzer DNA kit (Agilent Technologies, San Diego CA). Cluster generation and paired-end sequencing was performed according to the manufacturer's protocols (Illumina) at the Texas A&M AgriLife Genomics and Bioinformatics Center (College

Station, TX). In total, we performed 21 lanes of 75 bp paired-end and 2 lanes of 75 bp single-end sequencing using the Illumina Genome Analyzer II.

Sequence Mapping

Quality filtering was performed to remove reads and terminal bases with poor quality scores using the trim function in the CLC Genomics Workbench 4 (CLC Bio, Aarhus, Denmark) with the following parameters: ambiguous limit, 2; ambiguous trim, yes; quality limit, 0.1; quality trim, yes; and, remove 3' nucleotide, no; remove 5' nucleotide, no. The CLC Genomics Workbench Reference Mapping function was used to assemble the trimmed reads to the bosTau4.0 reference assembly using the following parameters: similarity score = 0.8; and, length fraction = 0.5. Paired-end reads were mapped using an insert range of 180-bp to 500-bp and reads matching multiple locations in the reference genome were placed using the random setting.

We created a *de novo* assembly of all reads that did not align to the reference genome (including ChrMt and ChrUn) with CLC Genomics' *De Novo* Assembly tool using the following parameters: similarity = 0.8; length fraction = 0.5; insertion cost = 3; deletion cost = 3; mismatch cost = 3; minimum paired distance = 180 bp; and, maximum paired distance = 500 bp. The resulting contigs were further analyzed by BLAST mapping to all complete genomes and chromosomes from RefSeq.

Variant Detection

We used the SNP detection function in CLC Genomics Workbench, based on the neighborhood quality standard (NQS) algorithm, using the following parameters: minimum coverage = 5; minimum central base quality = 30; average base quality over a window length of 11 nucleotides = 15; and, minimum allele frequency = 35%. Next, we used the deletion and insertion polymorphism (DIP) function in CLC Genomics

Workbench using the following parameters: minimum coverage = 5; minimum allele frequency = 35%; and, maximum expected variations = 2. Variants were further filtered to remove any located within the pseudo autosomal region (PAR) or within 10 bp of another variant.

In order to identify CNVs by sequence read-depth, we first used the Control-FREE Copy number (FREEC) program to identify CNVs in the mapped sequence data [68, 69]. CNVs were detected using two methods in FREEC; comparative (Angus vs Nellore) and independent (i.e., sample vs reference assembly). The Angus was selected as the reference and CNVs in the Nellore were detected by comparisons of read-depths across the genome. The comparison of CNVs identified by sequencing and array comparative genomic hybridization (CGH) allowed for the optimization of the following FREEC's parameters: breakpoint threshold = -0.001; window length = 10,000 bp; and, step = 5,000 by comparing. The independent calling of CNVs used the following parameters: bosTau4.0 reference genome; breakpoint threshold = -0.002; window length = 10,000 bp; and, step = 5,000 bp.

Next, the comparison of the Nellore against the Angus sequence data allowed for the identification of CNVs using CNV-Seq [72]. For running the CNV-seq script, we used the following settings: --log2-threshold = 0.6, --p value = 0.00001, --bigger-window = 3, --genome-size = cattle chromosome sizes (bp). These settings allowed for different windows sizes to be selected based on the read-depths and chromosome lengths.

Finally, we used MrFAST as an independent method to identify CNVs in the Angus and Nellore genomes [64]. The reference genome was masked using RepeatMasker (<http://repeatmasker.org>) and Tandem Repeat Finder (TRF) [180]. The raw data from each cow was mapped to all possible positions in the genome. Finally,

mrCaNaVar was used to report the GC corrected absolute copy numbers. A command line script was used to summarize the copy numbers. Duplications and deletions were identified using the following criteria: duplication, at least seven windows with a minimum copy number of 3; and deletion, at least 5 windows with a copy number less than one.

Variant Confirmation

The minimum read-depth used for variant analysis was determined by overlapping all SNVs from the Angus and Nellore with the samples' genotypes from the 770K BovineHD SNP BeadChip. The raw BovineHD SNP data Angus and Nellore from samples (Nellore 1-4 and Angus 1-4) were provided, as a service, by GeneSeek (NeoGen Corp, Lincoln, NE). All probes from the BovineHD array were converted to the bosTau4.0 genome assembly using the liftover tool in UCSC genome browser. The SNPs were then overlapped with repetitive regions masked by RepeatMasker and all bovine sequencing variants within 25 bp flanking each SNP. Additionally, all SNPs within homopolymers were identified by examining the base immediately before and after all SNPs. Finally, all SNPs were overlapped with tandem repeats downloaded from the UCSC genome browser. The Angus SNP genotypes were converted from AB allelic information to actual bases using the strand information from each probe. Using custom databases in ANNOVAR, the annotated BovineHD SNPs were overlaid with SNV data from the Angus genome.

Informative variants, those where the Angus' genotype is either homozygously or heterozygously different from the reference, were filtered using several different sets of filters to compare the false discovery rates (FDR), heterozygous undercall, and false negative rates (FNRs) by sequencing (Table 3.1). The BovineHD probes were first

filtered to remove all SNPs within 25 bp of other sequencing variants, homopolymers, repetitive regions, and unknown bases. A less stringent filtering method removed SNP array genotypes within 25 bp of another sequencing variant or where the probe contains an unknown base, N. Using these criteria, the accuracy of the SNP calling settings were compared to minimum read-depths between 5-10X for both heterozygous and homozygous variants.

Table 3.1 Definitions of terms used for the evaluation of accuracy SNP identification

Type	Definition	Calculation
$H^{SNP}H^{Seq}$	Number of homozygous SNPs (H^{SNP}) correctly identified as homozygous by sequencing (H^{Seq})	Sum of all $H^{SNP}H^{Seq}$
$H^{SNP}He^{Seq}$	Number of homozygous SNPs incorrectly identified as heterozygous by sequencing (He^{Seq})	Sum of all $H^{SNP}He^{Seq}$
$He^{SNP}H^{Seq}$	Number of heterozygous SNPs (He^{SNP}) incorrectly identified as homozygous by sequencing	Sum of all $He^{SNP}H^{Seq}$
$He^{SNP}He^{Seq}$	Number of heterozygous SNPs correctly identified as heterozygous by sequencing	Sum of all $He^{SNP}He^{Seq}$
$H^{SNP}X^{Seq}$	Number of homozygous SNPs not identified by sequencing	Sum of all $H^{SNP}X^{Seq}$
$He^{SNP}X^{Seq}$	Number of heterozygous SNPs not identified by sequencing	Sum of all $He^{SNP}X^{Seq}$
FDR^{Hom}	Error rate of correctly identifying homozygous SNPs	$\frac{(H^{SNP}He^{Seq})}{H^{SNP}He^{Seq} + H^{SNP}H^{Seq}}$
Heterozygous Undercall	He^{SNP} that are incorrectly genotyped as H^{Seq}	$\frac{(He^{SNP}H^{Seq})}{He^{SNP}H^{Seq} + He^{SNP}He^{Seq}}$
FNR^{Hom}	Proportion of H^{SNPs} missed by sequencing	$\frac{(H^{SNP}X^{Seq})}{H^{SNP}X^{Seq} + H^{SNP}He^{Seq} + H^{SNP}H^{Seq}}$

Table 3.1 Continued

Type	Definition	Calculation
FNR^{Het}	Proportion of He^{SNPs} missed by sequencing	$\frac{(He^{SNP}X^{Seq})}{He^{SNP}X^{Seq} + He^{SNP}He^{Seq} + He^{SNP}H^{Seq}}$
FNR^T	Proportion of ($H^{SNPs}+He^{SNPs}$) missed by sequencing	$\frac{(H^{SNP}X^{Seq} + He^{SNP}X^{Seq})}{H^{SNP} + He^{SNP}}$
GC	Genotypic concordance, the portion of BovineHD SNPs correctly genotyped by sequencing	$\frac{(H^{SNP}H^{Seq} + He^{SNP}He^{Seq})}{H^{SNP} + He^{SNP}}$

In addition to error predictions, we designed primers using Primer3Plus to amplify regions containing 34 SNVs in the Angus and Nellore (Appendix 3.2 [142]). Using standard PCR, regions containing SNVs were amplified. Amplicons were purified by gel purification (Qiagen) and Sanger sequenced by the Texas A&M DNA Technologies Lab. Sequences were aligned to the reference using both BLAT and ClustalW [181, 182]). All identified SNVs were compared to calls made from the next-generation sequencing data.

The CNVs identified by sequencing (all algorithms) were compared back to CNVs identified by aCGH to determine the amount of overlap and discrepancies between the variants. CNVs located within intergenic regions, terminal 1 Mb of chromosomes, and overlapping variants between the Angus and Nellore were removed from the comparison, as these would be less likely to have been identified by competitive hybridization of coding regions.

Genetic Variant Annotation and Analysis

We re-annotated the variant calls from CLC Genomics using both Galaxy (<http://galaxy.psu.edu/>) and ANNOVAR software programs [145, 183, 184]. The join and merge functions in Galaxy were used to annotate the SNVs and INDELS. We used these functions to compare the SNVs to all known SNVs in dbSNP. The ANNOVAR program used the Ensembl and RefSeq annotation databases to create an mRNA library, allowing for the determination of amino acid changes. The program was used to determine the locations of all variants within the genome. Variants identified as being unique to Angus and Nellore were combined to represent differences between the Angus and Nellore, while the total variants in each sample represent differences between the sample and the reference Hereford genome. The SNVs were divided into groups based on radical and conservative amino acid changes, where radical SNVs result in a difference in polarity or charge when the amino acid is changed while conservative SNVs cause no change in polarity or charge. Additionally, similar groups were created for variants representing differences between the Angus and Nellore. The biological functions of the genes affected by SNVs, INDELS, and CNVs were analyzed using the DAVID Functional Annotation Tool with the default settings [148, 149]. The resulting biological process terms were further grouped by similarities in function to determine enrichment for specific biological processes (Appendix 3.3). Statistical significance (p value) of enrichment in the defined groups was determined using Fisher's combined probability test with the p value created from the DAVID Functional Annotation Tool.

Genotyping for Known Mutations

A custom database was created with 88 known variants associated with phenotypes including production and disease. Using ANNOVAR, all variants in both the Angus and Nellore were compared to the custom phenotypic variant database. Overlapping variants were checked to determine whether the samples were carriers of the traits.

Evolutionary Analysis

Predicted coding sequences (CDS) for the cattle were created by incorporating all homozygous SNVs identified into the bovine genome assembly (bosTau4.0). Protein coding sequences were aligned using the clustalw software [185]. We used the PAML4.0 software package to determine the non-synonymous substitution rate (dN) and synonymous substitution rate (dS) of each protein [186]. These analyses involve pairwise comparison of two models and apply likelihood ratio test (LRT) to evaluate the significance. Two types of tests have been used in our study: a branch-specific test and a site-specific model. The branch-specific test (free ratio model vs. one ratio model) identifies positive selection signatures on specific evolutionary lineages in the phylogenetic tree. The site specific model compares the positive selection model with a nearly neutral model to identify the particular portion regions potentially positively selected. The statistical significance threshold was set to $p < 0.05$.

RESULTS

Genome Sequencing

The sequencing reactions yielded 1,402,540,397 sequence reads, totaling 98.3 Gb of DNA. The reads from the Nellore and Angus genomes were independently mapped to approximately 91% of the assembled autosomes and X chromosome of the

reference Hereford genome (bosTau4.0). Mapping resulted in the 18.6X and 17X read-depths of the Angus and Nellore genomes, respectively, with individual chromosome depths ranging from 14.3X to 31.6X (Table 3.2, Figure 3.1). However, the depth of ensembl annotated exons was much lower, with only 33% and 28% having at least 5X coverage in the Angus and Nellore, respectively. Approximately 76% of the reads from each animal mapped to unique positions, which is greater than previous reports of next-generation sequencing in cattle [129]. Reads not mapped to the assembled chromosomes were then mapped to the unassembled chromosomes (ChrUn), where approximately 40% of the contigs were covered in both genomes. Additionally, we generated 2,802X and 2,316X coverage of the Angus and Nellore mitochondrial genomes, respectively. The *de novo* assembly of 4,358,553 of the 11,241,919 unmapped Nellore reads resulted in 77,636 contigs with a minimum length and N50 of 200 bp and 336 bp, respectively. (Table 3.2)

Table 3.2 Overview of whole-genome sequencing data

Sample	Method	Lanes	Reads	Bases Mapped (Gb)	Average Depth of Coverage	Average Exome Depth of Coverage	% of Reference Mapped
Angus	75 SE	1	28,839,216	50,861,502,387	18.6 X	6.3 X	91%
Angus	75 PE	10	692,053,232				
Nellore	75 SE	1	20,783,003	47,471,311,170	17.3 X	5.5 X	91%
Nellore	75 PE	11	660,864,946				

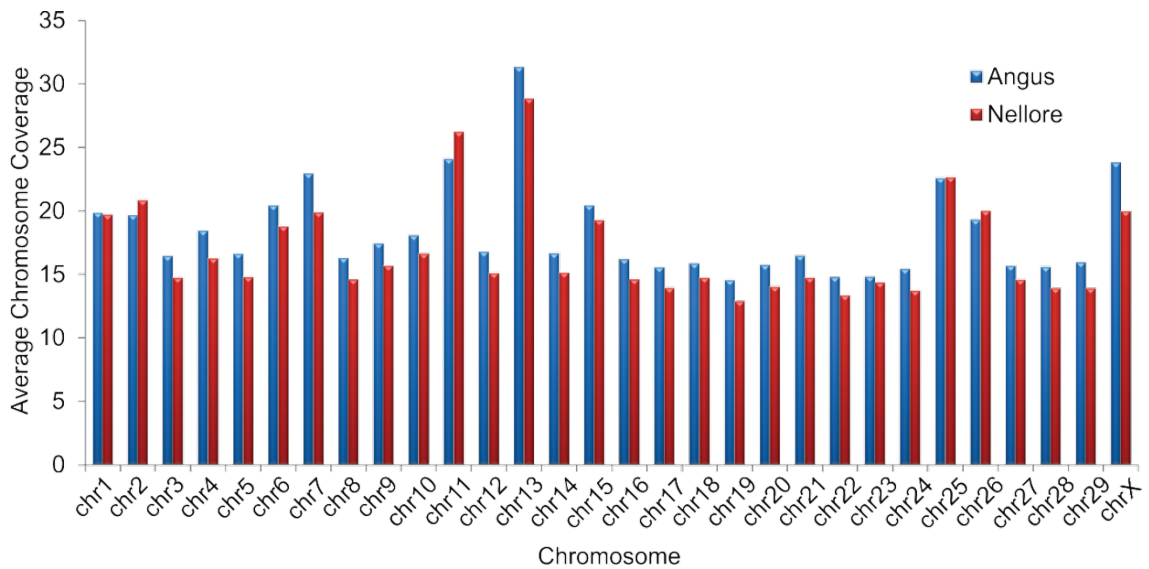


Figure 3.1 Read-depths across assembled chromosomes

Variant Identification and Annotation

The BovineHD SNP array was used to determine the FDR, FNR, and heterozygous undercall of the SNV calling parameters. The use of strict filtering of the BovineHD SNP genotypes, including the removal of SNPs overlapping repeat masked regions and homopolymers, was not found to improve the calling rates. With only 73,196 SNPs overlapping with strict filtering, the FDR^{Hom} was 1.9%. Therefore, the number of SNPs overlapping SNVs was increased by relaxing the filtering criteria to only remove probes with: an unknown base (N), within 25 bp of other sequencing variants, within the PAR, and shared genotypes with the reference genome. Increasing the minimum fold coverage required for SNV detection led to a significant loss of variants, without improving the FDR. Thus, a minimum of 5X coverage was used for the identification of SNVs in each animal. At 5X coverage, 99% of the overlapping array and sequencing genotypes were concurrent (Table 3.3). Additionally, the stringent minimum

allele frequency (35%) combined with an average coverage below the recommended 30X resulted in a heterozygous under-call rate of 23% [70]. Based on our average coverage between 15 and 20X, it would be expected that approximately 65% of the BovineHD SNP positions could be called by sequencing [70]. Additionally, the estimated 65% and 45% of callable genomic and coding regions appear to correspond well with our data. Finally, Sanger sequencing confirmed 33 of 34 randomly (97%) selected SNVs.

Table 3.3 Evaluation of accuracy of SNV identification using relaxed filtering

Type	5X	6X	7X	8X	9X	10X
$H^{SNP}H^{Seq}$	75,444	64,753	55,846	48,375	42,307	37,198
$H^{SNP}He^{Seq}$	821	722	682	632	557	528
$He^{SNP}H^{Seq}$	14,134	11,994	8,826	7,057	6,009	4,702
$He^{SNP}He^{Seq}$	48,204	42,019	39,522	35,738	31,690	29,371
$H^{SNP}X^{Seq}$	29,731	42,562	54,637	63,877	70,993	77,409
$He^{SNP}X^{Seq}$	63,732	72,057	77,722	83,275	88,371	91,997
FDR^{Hom}	1.08%	1.10%	1.21%	1.29%	1.30%	1.40%
Heterozygous Undercall	22.67%	22.21%	18.26%	16.49%	15.94%	13.80%
FNR^{Hom}	36.77%	45.73%	53.19%	59.45%	64.54%	68.82%
FNR^{Het}	61.76%	66.67%	68.65%	71.65%	74.86%	76.70%
FNR^T	43.51%	51.30%	57.27%	62.59%	67.17%	70.74%
GC	53.28%	45.61%	40.20%	35.20%	30.84%	27.60%

Genomic variant maps, including SNVs, INDELS, and CNVs were generated by comparisons made with the reference bovine assembly and by identifying variants between each genome. Using uniquely mapped reads with a coverage ≥ 5 and a quality score >30 , we identified 3,925,205 SNVs in the Angus genome and 6,931,681 SNVs in the Nellore genome (Table 3.4). Comparison of identified SNVs in each genome to the bovine SNP database (dbSNP, <http://www.ncbi.nlm.nih.gov/projects/SNP/>) revealed that 2,007,578 and 5,077,454 SNVs were novel in the Angus and Nellore genomes, respectively (Table 3.4). Of the SNVs identified, 6,454 were complex (i.e., heterozygous with both alleles different from the reference genome) in the Angus and Nellore genomes. Additionally, we identified 14 and 23 SNVs in the Angus and Nellore mitochondrial genomes, respectively. The comparison of SNVs in the Angus and Nellore revealed 7,255,802 variants between them (Table 3.5). The complete annotation analyses, with minimum read-depths from 5 to 10X, are listed in Appendix 3.4.

Table 3.4 Annotation of SNVs identified in comparison to the reference Hereford genome

	Total		Homozygous		Heterozygous		Novel		Ensembl Genes	
	Nellore	Angus	Nellore	Angus	Nellore	Angus	Nellore	Angus	Nellore	Angus
All SNPs	6,931,681	3,925,205	3,408,786	1,738,841	3,522,895	2,186,364	5,077,454	2,007,583	21,033	20,690
Intergenic	5,106,262	2,902,109	2,485,650	1,270,638	2,620,612	1,631,471	3,762,819	1,506,217	17,146	16,660
Intergenic (Upstream w/in 1 kb)	39,697	25,301	19,844	11,294	19,853	14,007	27,836	12,957	11,531	8,151
Intergenic (Downstream w/in 1 kb)	41,320	26,602	20,385	11,415	20,935	15,187	29,205	14,020	11,533	8,312
Intergenic (Up/Down w/in 1 kb)	656	393	325	187	331	206	464	191	292	192
Genic	1,743,746	970,800	882,582	445,307	861,164	525,493	1,257,130	474,198	17,677	16,619
Intron	1,700,197	937,533	862,003	431,475	838,194	506,058	1,226,395	454,790	15,594	14,822
Non-Coding Exon	2,319	1,740	968	629	1,351	1,111	1,792	1,171	907	628
5' UTR	1,262	929	635	456	627	473	854	505	729	492
3' UTR	9,456	6,779	4,782	3,054	4,674	3,725	6,612	3,655	3,487	2,474
Intron Splice Site	367	274	173	128	194	146	245	154	320	242
Exon Splice Site	647	429	298	192	349	237	436	217	556	358
Coding Exon	29,498	23,116	13,723	9,373	15,775	13,743	20,796	13,706	8,610	6,448
Synonymous	15,454	10,591	7,604	4,698	7,850	5,893	10,709	5,796	6,698	4,538
Non-Synonymous	14,321	12,600	6,301	4,771	8,020	7,829	10,214	7,874	5,182	4,126
Radical	9,258	8,172	4,038	3,066	5,220	5,106	6,622	5,180	3,974	3,153
Conservative	5,063	4,428	2,263	1,705	2,800	2,723	3,592	2,694	2,822	2,250
Stop Gain	355	334	109	87	246	247	297	239	299	270
Stop Loss	15	20	7	9	8	11	12	14	15	20

Table 3.5 Annotation of SNVs identified between the Angus and Nellore genomes

	Total	Homozygous	Heterozygous	Novel	Ensembl Genes
All SNPs	7,255,802	3,011,415	4,244,387	5,557,121	21,153
Intergenic	5,320,761	2,172,514	3,148,247	4,104,740	17,317
Intergenic (Upstream w/in 1 kb)	41,990	17,763	24,227	30,684	11,674
Intergenic (Downstream w/in 1 kb)	43,650	18,366	25,284	32,066	13,165
Intergenic (Up/Down w/in 1 kb)	738	329	409	512	348
Genic	1,848,663	802,443	1,046,220	1,389,119	17,751
Intron	1,799,630	782,389	1,017,241	1,354,046	15,996
Non-Coding Exon	2,554	850	1,704	2,076	453
5' UTR	1,360	593	767	950	343
3' UTR	10,691	4,767	5,924	7,458	4,301
Intron Splice Site	383	142	241	288	447
Exon Splice Site	677	265	412	490	715
Coding Exon	33,368	13,437	19,931	23,811	10,281
Synonymous	17,486	7,783	9,703	12,057	7,583
Non-Synonymous	16,098	5,809	10,289	11,862	6,173
Radical	10,438	3,711	6,727	7,760	4,757
Conservative	5,660	2,098	3,562	4,102	3,401
Stop Gain	442	100	342	368	384
Stop Loss	19	10	9	14	19

The identification of small (1-8bp) INDELs in the Angus and Nellore genomes revealed over 250,000 variants. The majority of the variants affected non-coding regions of the genome (introns and intergenic regions) (Table 3.6). The INDELs were predominately single base mutations, with very few 8 bp INDELs being identified (Figure 3.2). Unique INDELs identified only in one sample represent the genetic variation between the Angus and Nellore (Table 3.7).

Table 3.6 Annotation of INDELs identified in comparison to the reference Hereford

	Total		Homozygous		Heterozygous	
	Nellore	Angus	Nellore	Angus	Nellore	Angus
INDELs (1 bp - 8 bp)	159,794	117,813	62,609	47,964	97,185	69,849
Intergenic	118,640	87,335	46,679	35,797	71,961	51,538
Intergenic (Upstream w/in 1 kb)	813	697	326	290	487	407
Intergenic (Downstream w/in 1 kb)	820	661	293	239	527	422
Intergenic (Up/Down w/in 1 kb)	14	12	4	3	10	9
Genic	39,507	29,108	15,307	11,635	24,200	17,473
Intron	39,128	28,682	15,201	11,538	23,927	17,144
Non-Coding Exon	20	15	4	7	16	8
5' UTR	12	22	5	6	7	16
3' UTR	138	141	55	38	83	103
Intron Splice Site	11	13	5	4	6	9
Exon Splice Site	16	23	7	10	9	13
Coding Exon	182	212	30	32	152	180
Frameshift Deletion	68	74	12	19	56	55
Frameshift Insertion	109	137	20	20	89	117
Frameshift Substitution	0	1	0	0	0	1
Non-Frameshift Insertion	4	4	1	2	3	2
Non-Frameshift Substitution	13	13	4	1	9	12
Stop-Gain	4	6	0	0	4	6

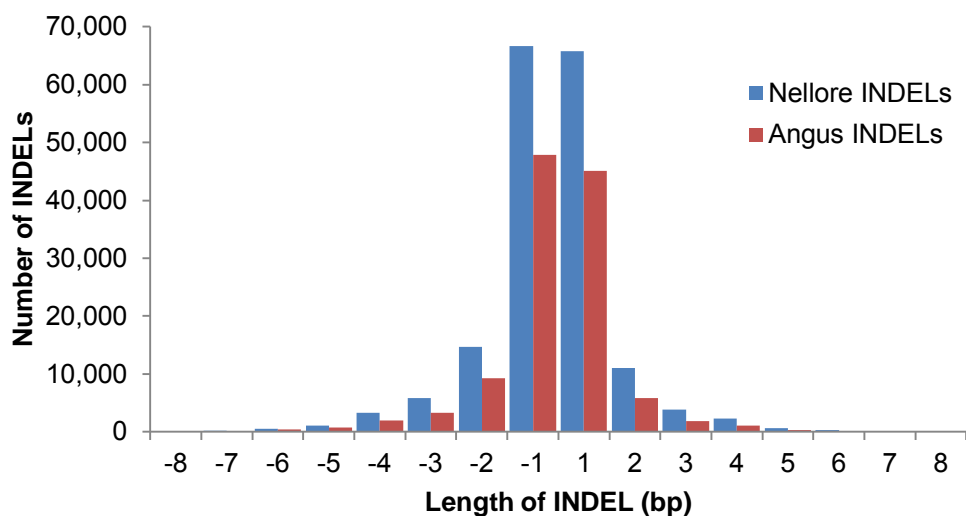


Figure 3.2 Distribution of INDELs by variant length

Table 3.7 Annotation of INDELs between the Angus and Nellore genomes

	Total	Homozygous	Heterozygous
INDELs (1 bp - 8 bp)	228,448	77,864	150,584
Intergenic	168,559	57,718	110,841
Intergenic (Upstream w/in 1 kb)	1,217	426	791
Intergenic (Downstream w/in 1 kb)	1,242	396	846
Intergenic (Up/Down w/in 1 kb)	26	7	19
Genic	57,404	19,317	38,087
Intron	56,669	19,151	37,518
Non-Coding Exon	31	8	23
5' UTR	28	8	20
3' UTR	260	84	176
Intron Splice Site	22	9	13
Exon Splice Site	29	9	20
Coding Exon	365	48	317
Frameshift Deletion	126	24	102
Frameshift Insertion	225	27	198
Frameshift Substitution	1	0	1
Non-Frameshift Insertion	6	1	5
Non-Frameshift Substitution	26	5	21
Stop-Gain	10	0	10

Copy number variants (CNVs) were identified using four read-depth (RD) based methods, consisting of independent and comparative RD measurements within each genome. Method-1 consisted of an independent CNV analysis of read-depths in the Angus and Nellore genomes using FREEC. Due to the errors associated with independent CNV detection from GC corrected read-depths, we chose a large window size and removed all variants within 1Mb of the chromosomal ends. These stringencies resulted in equal numbers of large CNVs in the Angus and Nellore, affecting over 200Mb of each genome (Table 3.8). The CNVs were present in over 1,700 genes and over 200 intergenic regions. Comparison of the 796 CNVs resulted in 135 CNVs between the Angus and Nellore. The large amount of CNVs sharing between the Angus and Nellore genomes suggests that many variants may in fact be segmental duplications. The Angus and Nellore genomes were further analyzed for CNVs by a comparison of read-depths using FREEC (Method-2) (Table 3.8). The resulting CNVs were further filtered as in method-1. The resulting 701 CNVs affected 18.5 Mb. The third method uses CNV-Seq to perform comparative analysis of CNVs by read-depths (Method-3, Table 3.8). The windows in this method were much smaller, but with more stringency on levels of differences required to identify a CNV. The resulting 330 CNVs were located in both genic and intergenic regions, while affecting 4.4 Mb of sequence. Finally, Method-4 utilized the MrFAST algorithm to perform independent CNV analyses on repeat masked genome sequences (Method-4, Table 3.8). The masking and differences in window sizes resulted in nearly equivalent numbers of CNVs in both genomes, with approximately 70% of the total CNVs being found in both samples.

In a comparison between the programs, we found that approximately 30% of CNVs were identified using both independent methods (Method-1 and -4), while only

23% were shared using comparative methods (Method-2 and -3). The CNVs from each method were compared to those identified by the exome CGH array (comparative methods) and BovineHD SNP array (independent methods). See Chapter V for CNV analysis using BovineHD SNP array. Due to differences in probe spacing and resolutions of methods, CGH CNVs were filtered to remove any below the minimum resolution of the sequencing method. Additionally, intergenic CNVs by sequencing were removed, as these regions were not analyzed by the exome array. The comparison of the CNVs indicated that more CNVs were concurrent between the comparative methods (Method-2:42.6%, and Method-3:33.3%) than in the independent methods (Method-1:20%, and Method-4:30%). We also found that the majority of independently identified CNVs were located in segmental duplications (72-81%), while less of the comparative CNVs were located in SDs (46-65%). In all cases, as was seen with the CGH analyses, larger CNVs were more likely to be located within SDs.

Table 3.8 Independent and comparative CNV analyses using sequence read-depth

Method 1			
Sample	# CNVs (Genes)	Size Range	Affected BP
Angus	398 (1707)	15 - 11,475 kb	224.5 Mb
Nellore	398 (1611)	15 - 11,300 kb	208.4 Mb
Angus vs Nellore	135 (1030)	15 - 11,300 kb	72.3 Mb
Method 2			
Sample	# CNVs (Genes)	Size Range	Affected BP
Angus	-	-	-
Nellore	-	-	-
Angus vs Nellore	701 (506)	10-815 kb	18.6 Mb

Table 3.8 Continued

Method 3			
Sample	# CNVs (Genes)	Size Range	Affected BP
Angus	-	-	-
Nellore	-	-	-
Angus vs Nellore	330 (134)	5.1-168.5 kb	4.4 Mb
Method 4			
Sample	# CNVs (Genes)	Size Range	Affected BP
Angus	541 (341)	4 - 131.6 kb	13.7 Mb
Nellore	576 (381)	4 - 131.6 kb	14.1 Mb
Angus vs Nellore	316 (190)	4 - 91.5 kb	4.2 Mb

Functional Analysis of Variants

Analysis of the coding portion of the genome revealed that approximately 67.6% of the protein-coding genes were identical between the Angus and Nellore genomes, while 7% and 25% had similar (e.g., containing only conservative amino acid changes) and divergent (e.g., containing radical amino acid changes) amino acid sequences, respectively. Coding variants (i.e., nonsynonymous) in the Nellore were primarily enriched in immunity and defense ($p = 5.1 \times 10^{-5}$), signal transduction ($p = 1.6 \times 10^{-6}$) and sensory perception ($p = 6.4 \times 10^{-9}$) pathways (Figure 3.3). Conversely, coding variation between the Angus and reference Hereford genome was primarily enriched in signal transduction ($p = 5.8 \times 10^{-5}$), sensory perception ($p = 1.1 \times 10^{-8}$) and immunity and defense ($p = 6.5 \times 10^{-2}$). CNVs were primarily enriched in immunity and defense ($p = 4.1 \times 10^{-8}$) pathways, however, INDELs were not significantly enriched ($P < 0.05$) in any biological process. Further comparison of radical nsSNVs indicated the Nellore sample possessed a significantly higher enrichment for immunity and defense

processes when compared to both *Bos taurus* samples (Figure 3.4). However, the enrichment for immunity and defense processes affected by conservative nsSNVs were not significantly different between the samples.

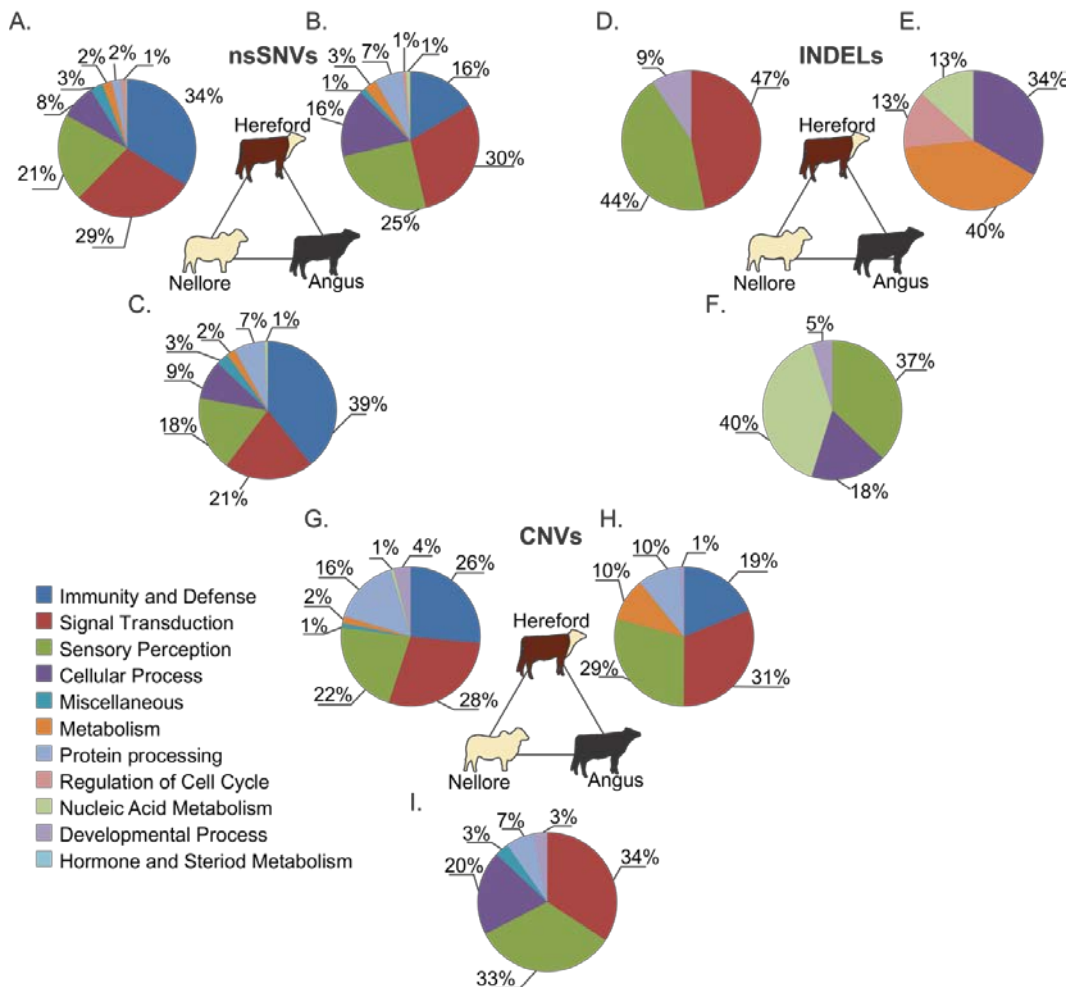


Figure 3.3 Biological process (BP) analysis of SNVs, INDELs, and CNVs. BP enrichment of genes with nsSNVs between (A.) Nellore and Hereford, (B.) Angus and Hereford, and (C.) Nellore and Angus. BP enrichment of genes with coding INDELs between (D.) Nellore and Hereford, (E.) Angus and Hereford, and (F.) Nellore and Angus. BP enrichment of genes with CNVs between (G.) Nellore and Hereford, (H.) Angus and Hereford, and (I.) Nellore and Angus.

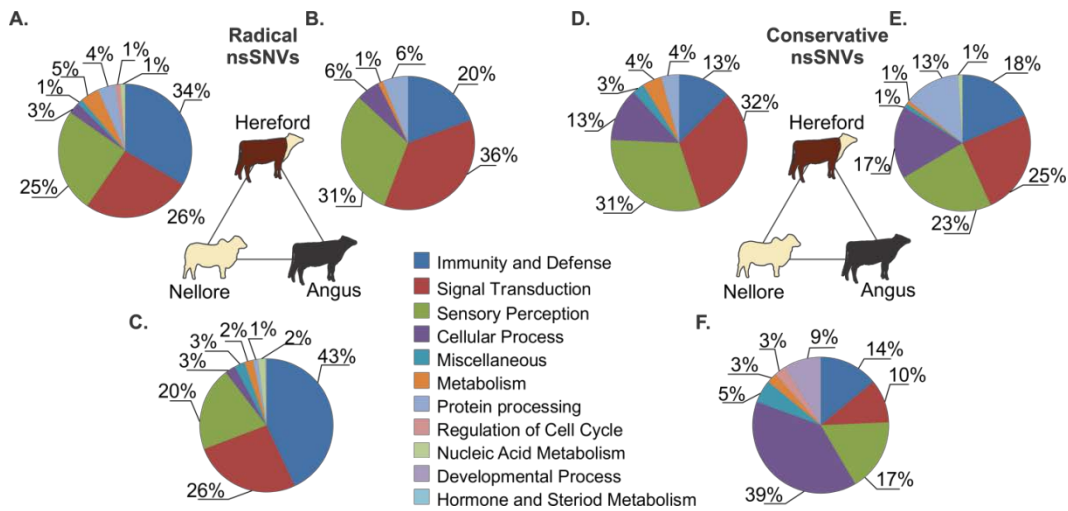


Figure 3.4 Biological process (BP) analysis of radical and conservative nsSNVs. BP enrichment of genes with radical nsSNVs between (A.) Nellore and Hereford, (B.) Angus and Hereford, and (C.) Nellore and Angus. BP enrichment of genes with conservative nsSNVs between (D.) Nellore and Hereford, (E.) Angus and Hereford, and (F.) Nellore and Angus

Genotyping for Known Mutations

In order to further understand the biological significance of identified variants, a database of known variants with association to specific phenotypes was created. The comparison of all variant locations with the sequencing data using the Pileup function of SAMTOOLS revealed that approximately 30% of causal mutation positions could be accurately genotyped using a minimum coverage of 5X. Of the genotypes that could be determined, the Angus was found to possess a heterozygous mutation associated with coat color spotting in Holstein and Simmental cattle [187]. Further attempts to genotype variants with read-depths below 5X were carried out using the data from the SAMTOOLS pileup file. This resulted in the identification of a heterozygous dinucleotide variant in the *DGAT1* gene of the Nellore. The *DGAT1* mutation is known to affect milk fat and protein production [188] (Table 3.9). The samples did not contain any of the

disease causing mutations. Furthermore, no nsSNVs were identified in regions previously found to be under selection within *Bos indicus* and other cattle [102, 151].

Table 3.9 Genotyping for known casual and associated mutations for diseases and traits in cattle

PMID	Chr	Coordinate	Gene	Phenotype	Associated Genotype	Angus Genotype	Nellore Genotype
8486364	chr1	70,332,664	<i>UMPS</i>	Deficiency of uridine monophosphate synthase	T/T	ND	ND
12047224	chr1	78,468,640	<i>CLDN16</i>	Renal dysplasia	56,002 bp Del	Wt	Wt
10995564	chr1	78,469,113	<i>CLDN16</i>	Renal dysplasia	36,910 bp Del	Wt	Wt
1384046	chr1	146,773,988	<i>ITBG2</i>	Bovine leukocyte adhesion deficiency	G/G	A/A	A/A
1384046	chr1	146,773,988	<i>ITGB2</i>	Leukocyte adhesion deficiency, type I	G/G	A/A	A/A
15776436	chr10	63,509,608	<i>FBN1</i>	Marfan syndrome	A/A	ND	ND
19714378	chr11	15,379,543	<i>SPAST</i>	Spinal dysmyelination	A/A	ND	ND
2813370	chr11	104,500,479	<i>ASS1</i>	Citrullinaemia	T/T	ND	ND
17033029	chr11	107,166,537	<i>PAEP</i>	Beta-lactoglobulin, aberrant low expression	A/A	ND	ND
16935476	chr12	52,552,849	<i>CLN5</i>	Neuronal ceroid lipofuscinosis, 5	GG/GG	G/G	G/G
11827942	chr14	445,086	<i>DGAT1</i>	Milk fat, protein%	AA/AA	ND	GC/AA
3472203	chr14	7,873,527	<i>TG</i>	Goitre, familial	A/A	ND	ND
19398771	chr15	20,782,835	<i>BCO2</i>	Yellow fat	A/A A/G	G/G	G/G
16963222	chr15	77,142,483	<i>LRP4</i>	Syndactyly (mule foot)	T/T	G/G	ND
16859890	chr15	77,150,869	<i>LRP4</i>	Syndactyly (mule foot)	AT/AT	ND	ND
16714095	chr17	21,442,628	<i>SLC39A4</i>	Acrodermatitis enteropathica	A/A	ND	ND
23029151	chr17	56,496,012	<i>KDM2B</i>	Lethal multi-organ developmental dysplasia	A/A	ND	ND
19374945	chr18	9,764,190	<i>WFDC1</i>	Multiple ocular defects	ins C/C	ND	ND
11467827	chr18	13,777,466	<i>MC1R</i>	Coat colour, extension in <i>Bos taurus</i> (cattle)	C/C T/C	ND	ND
8661706	chr18	13,777,481	<i>MC1R</i>	Coat colour, extension	Del G/G	ND	G/G
10425233	chr18	50,234,296	<i>BCKDHA</i>	Maple syrup urine disease	T/T	ND	C/C
10425233	chr18	50,243,638	<i>BCKDHA</i>	Maple syrup urine disease	T/T	ND	ND
19016676	chr18	52,796,321	<i>PPP1R13L</i>	Cardiomyopathy and woolly haircoat syndrome	Dup CGCCTGT	WtWt	WtWt
20923700	chr18	52,901,555	<i>OPA3</i>	Cardiomyopathy, dilated	A/A	G/G	ND
21152099	chr18	64,409,296	<i>MIMT1</i>	Abortion and stillbirth	106kb Del	Wt	Wt
12481987	chr19	26,860,160	<i>CHRNE</i>	Myasthenic syndrome, congenital	20 bp Del	WtWt	WtWt
8621763	chr19	45,488,394	<i>SLC4A1</i>	Spherocytosis	A/A	G/G	G/G

Table 3.9 Continued

PMID	Chr	Coordinate	Gene	Phenotype	Associated Genotype	Angus Genotype	Nellore Genotype
19779552	chr19	48,636,985	<i>MRC2</i>	Tail, crooked	*/*	AG/AG	ND
19524387	chr19	49,660,133	<i>GH1</i>	Dwarfism, growth-hormone deficiency	T/T	ND	ND
1072372	chr19	54,001,954	<i>GAA</i>	Glycogen storage disease II	**/**	ND	ND
1072372	chr19	54,005,930	<i>GAA</i>	Glycogen storage disease II	A/A	ND	ND
1072372	chr19	54,009,386	<i>GAA</i>	Glycogen storage disease II	**/**	ND	ND
9356471	chr2	6,533,052	<i>MSTN</i>	Muscular hypertrophy (double muscling)	A/C A/A	C/C	ND
11105210	chr2	6,535,208	<i>MSTN</i>	Muscular hypertrophy (double muscling)	T/T	ND	C/C
9288100	chr2	6,537,448	<i>MSTN</i>	Muscular hypertrophy (double muscling)	Hom 11 bp Del	Wt/Wt	Wt/Wt
9356471	chr2	6,537,568	<i>MSTN</i>	Muscular hypertrophy (double muscling)	A/A	ND	ND
23152852	chr2	51,537,189	<i>ZEB2</i>	Polled and multisystemic syndrome	3,740,690 bp Del	Wt	Wt
17952705	chr21	20,191,991	<i>ACAN</i>	Dwarfism, Dexter	GGCA Ins	Wt/Wt	ND
10357109	chr21	20,197,816	<i>ACAN</i>	Ehlers-Danlos syndrome	A/A	ND	ND
22952632	chr21	20,537,018	<i>FANCI</i>	Brachyspina	Del 3,329bp	Wt	Wt
22174915	chr22	32,364,257	<i>MITF</i>	Dominant white with bilateral deafness	A/A C/A	ND	ND
22486495	chr22	32,386,957	<i>MITF</i>	Coat colour, white spotting	T/T	A/T	A/A
22715415	chr22	52,504,015	<i>COL7A1</i>	Epidermolysis bullosa, dystrophic	T/T	C/C	C/C
21255426	chr23	14,398,230	<i>MOCS1</i>	Arachnomelia, BTA23	Hom Del CA	ND	ND
9784594	chr24	59,105,614	<i>FECH</i>	Protoporphyrria	A/A	ND	ND
17420465	chr24	64,249,133	<i>KDSR</i>	Spinal muscular atrophy	T/T	C/C	C/C
18344998	chr25	27,746,956	<i>ATP2A1</i>	Congenital muscular dystonia 1	A/A	G/G	ND
23046865	chr25	27,752,641	<i>ATP2A1</i>	Pseudomytonia congenital	A/A A/C	C/C	ND
23046865	chr25	27,752,866	<i>ATP2A1</i>	Pseudomytonia congenital	A/A A/C	ND	ND
18786632	chr25	27,753,942	<i>ATP2A1</i>	Pseudomytonia congenital	T/T	C/C	C/C
17458708	chr26	32,389,019	<i>NAGLU</i>	Mucopolysaccharidosis IIIB	A/A	ND	ND
Unpublished	chr26	37,027,031	<i>GFRA1</i>	Forelimb-girdle muscular anomaly	A/A	G/G	G/G
16104386	chr27	17,619,151	<i>F11</i>	Factor XI deficiency	15 bp Ins	ND	ND
15566468	chr27	17,623,837	<i>F11</i>	Factor XI deficiency	Ins 76bp	Wt	Wt
10594238	chr28	6,945,386	<i>LYST</i>	Chediak-Higashi syndrome	C/C	T/T	T/T
Unpublished	chr29	300,100	<i>HEPHL1</i>	Hypotrichosis	A/A	ND	ND
14727143	chr29	6,536,066	<i>TYR</i>	Coat colour, albinism	C/C	ND	ND
18344998	chr29	25,553,812	<i>SLC6A5</i>	Congenital muscular dystonia 2	G/G	A/A	A/A
18039909	chr29	44,762,749	<i>RASGRP2</i>	Thrombopathia	G/G	A/A	A/A
8845714	chr29	44,775,330	<i>PYGM</i>	Glycogen storage disease V	A/A	ND	ND
16344554	chr3	46,229,040	<i>SLC35A3</i>	Complex vertebral malformation	A/A	ND	ND
22438830	chr3	101,805,195	<i>RNF11</i>	Dwarfism, proportionate, with inflammatory lesions	C/C	ND	T/T
21814570	chr4	29,020,047	<i>TWIST1</i>	Scurs, type 2	11bp Dup (cgggcccccgcg)	ND	ND

Table 3.9 Continued

PMID	Chr	Coordinate	Gene	Phenotype	Associated Genotype	Angus Genotype	Nellore Genotype
21526202	chr4	38,428,047	<i>MFN2</i>	Axonopathy	A/A	G/G	G/G
20507629	chr4	117,916,428	<i>SLC4A2</i>	Osteopetrosis	2,781 bp Del	Wt	Wt
10384045	chr5	20,591,714	<i>KITLG</i>	Coat colour, roan, white	A/A, A/C	ND	ND
15955091	chr5	30,270,142	<i>KRT5</i>	Epidermolysis bullosa	A/A	G/G	ND
20865119	chr5	61,881,761	<i>SUOX</i>	Arachnomelia, BTA5	insG	WtWt	ND
18408794	chr5	61,910,340	<i>PMEL</i>	Hypotrichosis with coat-colour dilution	3bp Del (Het)	WtWt	WtWt
17302792	chr5	61,910,352	<i>PMEL</i>	Coat colour, dilution in <i>Bos taurus</i>	A/A G/A	G/G	ND
18408794	chr5	61,918,184	<i>PMEL</i>	Hypotrichosis with coat-colour dilution	A/C	C/C	ND
10594236	chr6	23,801,171	<i>MANBA</i>	Mannosidosis, beta	A/A	G/G	ND
19887637	chr6	99,298,982	<i>PRKG2</i>	Dwarfism, Angus	A/A	G/G	G/G
12354143	chr6	107,836,847	<i>EVC2</i>	Chondrodysplasia	T/T	ND	ND
12354143	chr6	107,852,627	<i>EVC2</i>	Chondrodysplasia	Del CA Ins G	ND	CA/CA
10417273	chr7	1,885,655	<i>ADAMTS2</i>	Ehlers-Danlos syndrome, type VII (Dermatosparaxis)	17 bp Del	ND	ND
9491457	chr7	11,109,933	<i>MAN2B1</i>	Mannosidosis, alpha	A/A	G/G	G/G
9491457	chr7	11,111,242	<i>MAN2B1</i>	Mannosidosis, alpha	C/C	T/T	ND
11178872	chr7	62,833,910	<i>GLRA1</i>	Myoclonus	T/T	ND	G/G
18557975	chr8	33,475,363	<i>TYRP1</i>	Coat colour, brown	A/A	ND	ND
19456318	chrUn.0 04.16	354,250	<i>F8</i>	Haemophilia A	T/T	ND	ND
16827753	chrUn.0 04.288	9,628	<i>ASIP</i>	Coat colour, agouti	8404 bp ins	ND	ND
18344998	chrUn.0 04.3	1,131,276	<i>ABCA12</i>	Ichthyosis congenita	G/G	ND	ND
12466292	chrUn.0 04.3	1,737,477	<i>FMO3</i>	Trimethylaminuria	T/T	ND	ND
21410470	chrUn.0 04.31	197,320	<i>EDA</i>	Anhidrotic ectodermal dysplasia	19bp Del	ND	ND
22497423	chrUn.0 04.31	197,553	<i>EDA</i>	Anhidrotic ectodermal dysplasia	ins AGGG/AGGG	ND	ND
11591646	chrUn.0 04.31	437,834	<i>EDA</i>	Anhidrotic ectodermal dysplasia	~300bp Del	ND	ND
12021844	chrUn.0 04.31	448,839	<i>EDA</i>	Anhidrotic ectodermal dysplasia	G/G	ND	ND

Evolutionary Analysis

The SNVs in the Angus and Nellore samples were further analyzed using the dN/dS test on 11,521 RefSeq genes to determine if the identified coding SNVs were under positive selection in the *Bos indicus* or *Bos taurus* cow. The resulting branch-specific analysis revealed few genes (11) being identified as significantly under

selection in the *Bos indicus* vs *Bos taurus* cattle (Table 3.10). Using the site-specific model to identify genes under positive selection yielded only 11 and 27 genes with dN/dS scores greater than 1 in the Angus and Nellore, respectively (Appendix 3.5) (Table 3.11). The genome-wide average dN/dS for the Angus and Nellore cows were 0.10 and 0.19, respectively (Table 3.11). These data suggest a high level of sequence conservation between the subspecies.

Table 3.10 Genes with significant positive selection using the branch-specific model.

Gene	omega	Log Ratio Test (LRT)
<i>DNAH9</i>	0.56794	10.581282
<i>BLA-DQB</i>	0.12491	7.50576
<i>TRAF3IP1</i>	0.4753	7.630062
<i>THNSL2</i>	0.91139	8.36848
<i>CP</i>	0.15226	6.731074
<i>ATP2C1</i>	0.20061	6.780098
<i>ASPM</i>	0.23355	7.261786
<i>ERCC5</i>	0.39295	8.298822
<i>MGC134066</i>	0.37398	6.733036
<i>HERC2</i>	0.0001	6.877404
<i>CD46</i>	0.14997	6.748846

Table 3.11 Effects of selection using homozygous SNVs within RefSeq genes

Method	Genes	Angus	Nellore
Site-Specific	dN/dS >1	11	27
	Average dN	0.00051	0.000349
	Average dS	0.0034	0.00417
	Average dN/dS	0.1029	0.1925
	Identical	10,754	8,787
Branch-Specific	RefSeq	12,615	
	Analyzed	11,521	
	Significant	11	
	Identical	7,518	

DISCUSSION

Angus and Nellore cattle represent phenotypically diverse subspecies of cattle. The Angus breed was selected for meat production traits such as carcass quality, tenderness, and feed efficiency. The Nellore breed was selected for heat tolerance and disease resistance due to tropical origins. The hundreds of studies investigating these differences have identified over 5,000 QTLs but no causal mutations [146, 147, 176]. The recent advancements in next-generation sequencing technologies allowed for a burst in whole-genome sequencing of various species. Prior to this study, there have only been twelve published cattle sequences using next-generation technology. These descriptive studies have mostly focused on a single genome or have compared *Bos taurus* samples. The only study comparing a Nellore genome to *Bos taurus* samples focused primarily on CNVs and provided limited analysis of SNVs.

Our analysis of genetic variants across a single Angus and Nellore genome provides the first comparative insight into the genetic differences between *Bos indicus* and *Bos taurus* cattle. The massively parallel next-generation sequencing generated

98.3 Gb (billion bp) (50.8 Gb from the Angus and 47.4 Gb from the Nellore), resulting in 18.6X and 17.3X coverage of the Angus and Nellore genomes, respectively. The *de novo* assembly of unmapped reads created 77,636 contigs spanning 26 Mb in the Nellore genome. The combination of our previous analysis of CNVs in horses [155, 189], as well as those in humans and mice, demonstrate that large homozygous deletions are common in the genome of healthy individuals. Therefore, we suggest that some of the contigs represent novel sequence that is absent from the Hereford genome due to homozygous deletions.

The accuracy and ability to identify SNVs throughout the bovine genomes using the BovineHD SNP array indicated that our strict calling criteria increased the level of accuracy in identified SNVs, but reduced the accuracy of distinguishing between homozygous and heterozygous SNVs. Due to our stringency, we found that some heterozygous SNVs were incorrectly genotyped homozygous. The SNVs with the second allele frequency below our 35% threshold were excluded from the list. However, despite a fairly high heterozygous under-call rate, we were able to demonstrate a low FDR for genotyping homozygous SNVs. The high FNR, representing SNVs missed by sequencing, was found to be a result of stringent frequency thresholds, depth of coverage and the repetitive nature of many regions in the bovine genome. The fact that our exonic coverage was much lower than that of intergenic regions suggests that biases and variability in sequencing using the GAI platform exist. However, the introduction of Illumina's HiSeq technology may reduce these biases, resulting in a more even distribution of read coverage. Despite these limitations, we were able to identify over 10 million SNVs and 300,000 INDELS.

We identified CNVs by comparative and independent analyses of read-depth, including a duplication of cathelicidin 4 (*CATHL4*). An increase in copies of *CATHL4* is suspected to result in an increase in peptides and host resistance against pathogens, making this CNV an excellent candidate for further studies into the disease resistance of Nellore cattle. We confirmed CNVs by a comparison to known CNVs identified in the animals by a custom exon tiling array. The low percentage of overlap between the CNVs and those identified by CGH are due to differences in resolutions of the studies, existing duplications in the reference genome, and the known inconsistencies with depth of coverage tools [68]. This study, combined with human studies of copy number variation, demonstrates the need for the development of more accurate algorithms to detect CNVs by sequencing to minimize false discoveries and to identify all types of copy number variation.

Both SNVs and CNVs were shown to be enriched for biological processes of immunity and defense, signal transduction, and sensory perception. However, further investigation revealed that the Nellore possesses a threefold gain in enrichment of immunity and defense processes in radical SNVs over conservative SNVs, while levels remain similar in the Angus. Radical amino acid changes are thought to more drastically affected protein function than conservative changes due to the changes in charge and polarity [190, 191]. The increase in radical changes of immune genes could be a major factor in the increased disease resistance of the Nellore breed. However, the true link between the radical SNVs and immunity phenotypes will require further investigation.

Additionally, we created a custom bovine mutation database containing 88 known mutations that are causal and associated with traits in cattle. Using our database, we were able to genotype both animals for 88 known mutations, with the

majority causing severe diseases. However, the previously mentioned limitations of read coverage allowed us to only genotype 40 and 33 of the mutations in the Angus and Nellore genomes, respectively. The healthy animals did not carry any disease genotypes. Interestingly, the Angus was found to carry a heterozygous SNV that has been linked to spotting in other breeds [187]. However, the association to this specific intronic SNP was only found in Holstein and Simmental cattle. However, another non-spotted breed (Reggiana) did not reveal any association with this specific SNP. Therefore, it is unlikely that this SNP is actually predictive of the spotting in our Angus cow. Decreasing the depths required to call a SNV revealed that the Nellore possessed a heterozygous dinucleotide substitution in the diacylglycerol O-acyltransferase 1 (*DGAT1*) gene causing a K232A change in the protein. The K allele is responsible for a decrease in milk production but an increase in milk fat yield, percentage of fat, and percentage of protein [188, 192, 193]). Recent studies have shown that Holstein cattle have the highest frequency of the A allele, resulting in an increase in milk production. While our Nellore sample was heterozygous, studies show that only 1% of Nellore possess the A allele, while the remaining are homozygous for the K allele [194, 195].

The use of global dN/dS values for all genes by incorporating the Angus and Nellore SNVs into their genomes allowed for the identification of genes under differential selection in the breeds. The evolutionary analysis of coding variation revealed that the majority of coding genes remain identical or similar, with only 11 appearing to be under selection in one breed vs another. Additionally, when these genes along with nsSNVs were compared with known regions of selection in cattle, no overlap was identified, suggesting that while significant enrichment exists for immune processes, these nsSNVs are representative of a very small fraction of variants.

Therefore, many of the known quantitative traits may be controlled by other types of variation (i.e., CNVs) or by non-coding regions (i.e., regulatory elements).

Whole genome analysis revealed over 10 million SNVs, 300,000 INDELS, and 900 CNVs in the Angus and Nellore genomes. A combined functional analysis relying on biological processes and known mutations revealed enrichment for immunity processes and candidate immune related mutations in the Nellore sample. Despite the level of variation between the Angus and Nellore, dN/dS analyses indicate that very few genes possess significant divergence for positive selection between the breeds. These findings, combined with the lack of coding variants within known regions of selection, suggest that many quantitative traits may be controlled by other types of variation (i.e., CNVs) or by non-coding regions (i.e., regulatory elements). In order to better understand the underlying genetics of phenotypes in cattle, these data will need to be combined with epigenetic maps of the bovine genome.

CHAPTER IV

VARIATION AT REGULATORY ELEMENTS

INTRODUCTION

In the past 20 years, hundreds of studies have strived to identify the causal coding variants responsible for both Mendelian and complex traits in cattle. These studies have progressed from using single microsatellites to high-density SNP arrays to whole-genome sequencing. Despite these efforts, only approximately 88 causal mutations have been identified, with the majority leading to diseases or changes in coat color in *Bos taurus* cattle. However, the underlying causal genetic elements for quantitative traits such as fecundity, immunity, and production remain unknown. Recent attempts to narrow the over 5,000 known QTLs and regions of selection through the use of whole-genome sequencing have revealed a vast amount of unknown variation in cattle, further complicating the ability to identify causal variants. Despite the identification of more than 13 million SNPs and 2,000 CNVRs, many regions under selection lack known coding variants. These findings suggest that many traits in cattle may be controlled by variation at regulatory elements. Even with the existence of extensive gene annotations in the bovine genome, there are no publicly available regulatory element maps outside of 5' and 3' UTRs. The lack of epigenetic maps has limited the investigation of variants within cis-acting elements.

The identification and understanding of regulatory elements and genetic variation has proven to be a difficult task even within the well annotated human genome. The reason for the complexity is that regulation can occur at several stages including transcriptional, post-transcriptional, translational, and post-translational [196].

The transcriptional regulation of genes can be controlled by several types of cis-acting transcriptional DNA elements occurring around the promoter region or at distances of hundreds of kb or greater from the promoter. The promoters and proximal promoters (usually within 1 kb of the promoter) may have the greatest effect on transcription. However, the harder to identify intergenic and intronic regions such as enhancers, silencers, and locus control regions also have significant effects on the regulation of transcription. All of these regions can contain binding sites for specific transcription factors and, therefore, genetic variation within the elements could lead to alterations in gene transcription [197].

The recent advancements in human regulatory elements, including the ENCODE project, have demonstrated that several histone modifications and DNA methylation can be used not only to quantitatively identify active and repressed gene transcription, but also to identify genic and intergenic cis-acting regulatory elements. The usage of these modifications has resulted in the identification of nearly 400,000 enhancers and 70,292 promoter regions in the human genome [35]. Despite previous methods that used sequence conservation between species to identify regulatory elements, recent studies have shown that only 40 to 60% of regulatory elements and transcription factor binding sites are actually conserved between species [196, 198, 199]. The combination of these regions with structural variation and GWAS data suggests that many quantitative traits are controlled by regulatory element variation, while Mendelian traits are more likely to be controlled by coding variation [91-93]. Therefore, previous attempts that focused on coding regions for traits and complex diseases may have missed important underlying variation.

The importance of complex diseases and traits in cattle genomics led to the release of the bovine 50K SNP Beadchip, followed by the BovineHD SNP Beadchip. The utilization of these designs has led to thousands of cattle being genotyped, resulting in numerous regions thought to control traits, but very few causal variants. Based on the recent findings of non-coding variants causing traits and diseases in humans and other species, it is likely that many regions of selection and QTLs possess causal regulatory variants instead of coding variants. However, to date, there have been no studies in cattle to map regulatory elements using ChIP-seq or DNA methylation sequencing and subsequently overlay these regions with SNVs from whole-genome sequencing.

Therefore, we performed epigenetic profiling through a combination of chromatin immunoprecipitation (H3K4me3) and methyl-binding domain (5Mc) sequencing to identify putative regulatory elements in bovine WBCs. The combination of identified regulatory elements with SNVs from our whole-genome sequencing and previous studies provides the first glimpse into the extent and potential role of regulatory element variation on phenotypic differences between *Bos indicus* and *Bos taurus* cattle.

METHODS

Epigenetic Profiling

DNA modified with histone 3 lysine 4 tri-methylation (H3K4me3) was isolated from WBCs using an adapted version of a standard ChIP protocol for histone 3 lysine 4 tri-methylation (H3K4me3) [200]. Additionally, methylated CpG islands within WBCs were enriched using the MethylCollector Ultra kit (Active Motif, Carlsbad, Ca). All captured H3K4me3 and methylated DNA were used to create Illumina single-end sequencing libraries using the standard manufacturer's protocol. Each library was

sequenced on two 36 bp single-end lanes using an Illumina GAII at the Texas A&M AgriLife Genomics and Bioinformatics Center (College Station, TX).

All of our raw data, in FASTQ format, were mapped to the reference genome (bosTau4.0) using BWA within the Galaxy webserver using default settings [65]. Enriched regions were identified throughout the genome using MACS software with the following settings: band size, 150; Pvalue cutoff, 1e-05; MFOLD, 10; build model, create_model; and, futurefd, no. Annovar was used to annotate all resulting peaks for SNVs, INDELS, CNVs, ensembl genes, RefSeq genes, repetitive elements, QTLs, regions of selection (FST), conserved regions, and CpG islands.

Furthermore, all SNVs from our Angus and Nellore, along with those previously identified in a single Angus and Holstein, were annotated using the bovine ensembl database and our newly created bovine regulatory map [131]. The variant densities were determined for intergenic (IGR), intronic (INR), genic (introns, exons [except first or last exon], and 3'UTR), proximal promoter (PPR, [1kb upstream, 5' UTR, and exon 1]), transcriptional termination region (TTR, [1kb downstream, 3' UTR, and last exon]), coding exon (ER, [without first or last exon]), synonymous coding changes (ER(syn)), nonsynonymous coding changes (ER(NS)), regulatory elements (RE, [intergenic regions with H3K4]), and differentially modified CpG islands (DMC). The densities were compared between our single *Bos taurus* (Angus) and *Bos indicus* (Nellore) genomes, as well as a single sheep genome (Doan *et. al.* unpublished data). The OR of the Nellore vs the Angus was compared to the ORs of a sheep vs the Angus (Doan *et. al.* unpublished data) and the Holstein vs the Angus [131].

RESULTS

Identification of Regulatory Elements in the Bovine Genome

To identify regulatory elements in the bovine genome, we performed CHIP- and MBD-sequencing of captured DNA isolated from circulating lymphocytes. Regulatory elements were characterized as regions containing tri-methylated histone H3 lysine 4 (H3K4me3), a mark of active promoters and enhancers [93, 201]. Of the 17,005 modified regions identified, 62% and 60% overlapped CpG islands (CGIs) and conserved regions, respectively. However, only 48% of the regions overlapped both CGIs and conserved regions. The remaining 40% of regions without conservation were presumed to be regulatory elements unique to bovids. Regions enriched with H3K4me3 were annotated as being present at IGR, PPR, ER, INR, or TTR (Figure 4.1). While the majority of modified regions were located at PPRs (5,817) and IGRs (6,960), many were identified at ERs (431), INRs (3,388), and TTRs (246) (Appendix 4.1, Table 4.1). Further annotation revealed that 5,368 genes possess H3K4me3 at their PPR, suggesting active transcription. Additionally, many genes were affected by histone methylation at INRs (2,755), ERs (420), and TTRs (232). Interestingly, 7 genes have histone modifications at both PPR and TTR, while lacking it within the gene.

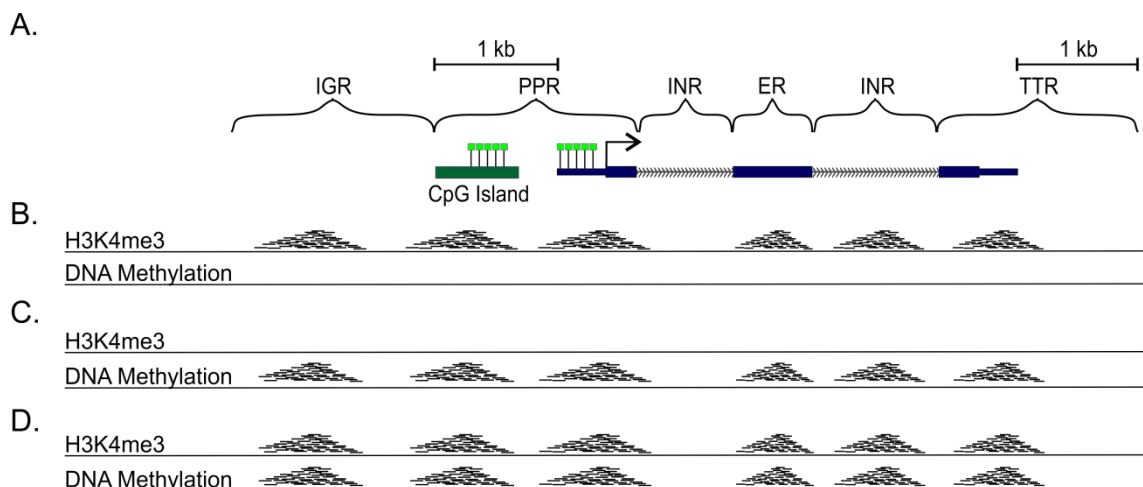


Figure 4.1 (A.) Definition of genomic regions used to identify regions (B.) with H3K4me3, (C.) DNA methylation, and (D.) with differential modifications

Table 4.1 Summary of epigenetic profiling for H3K4me3 and DNA methylation

Modification	Total	IGR	PPR	ER	INR	TTR
H3K4me3	17,005	6,960	5,811	436	3,388	250
DNA Methylation	44,486	19,576	2,871	10,340	8,618	2,826
DMC	718	200	318	38	111	25
Unique H3K4me3	15,087	5,872	5,459	373	3,060	210
Unique DNA	42,547	18,470	2,562	10,165	8,287	2,771

Given the functional roles of regions modified by H3K4me3, genetic variation at these regions could have a large effect on gene transcription. Therefore, we overlapped all genetic variants from sequencing and CGH analyses of the Angus and Nellore cow to determine the extent of the variation within IGRs, PPRs, ERs, INRs, and TTRs (Figure 4.2). We found of the 9,617 regions with 37,197 SNVs, the majority were located at IGR (3,642) and PPRs (3,501). Additionally, 1,948 INDELS were found to

affect 1,151 modified regions. While SNVs and INDELS would affect the binding of transcription factors, CNVs have the potential to duplicate or remove entire regulatory elements, leading to large changes in transcription. Overlapping of CNVs from our exome CGH experiments revealed 288 CNVs affecting 247 H3K4me3 regions, including 84 PPRs. Overall, despite the important functional role of H3K4me3, a large level of underlying genetic variation exists throughout many modified regions.

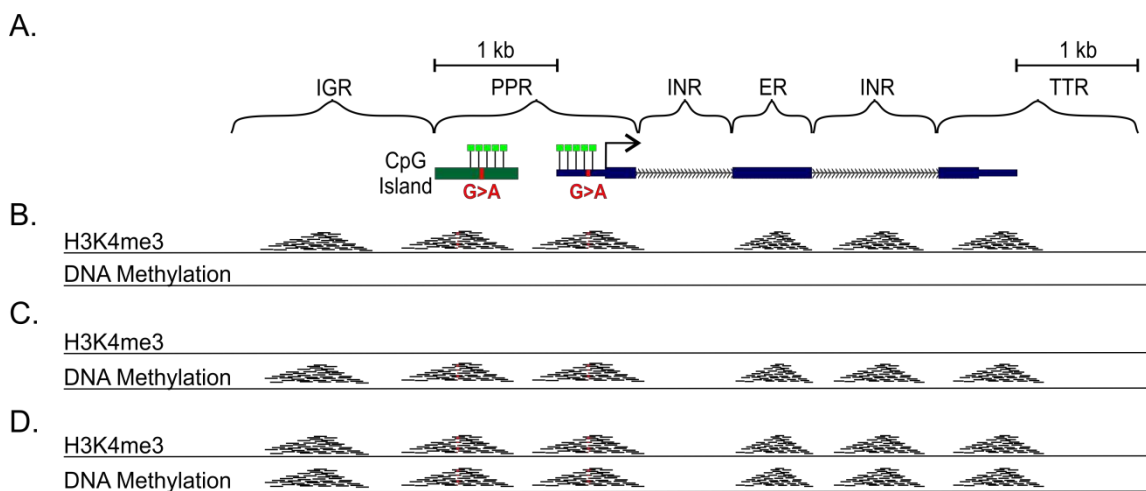


Figure 4.2 (A.) SNVs located within genomic regions (B.) with H3K4me3, (C.) DNA methylation and (D.) with differential modifications

In addition to active H3K4me3 histone methylation, enrichment of DNA methylation, an indicator of repression, was investigated throughout the bovine genome. The detection of peaks (regions) enriched for DNA methylation revealed 44,486 regions throughout the bovine genome. Approximately, 33% of DNA methylation regions occurred at CGIs, accounting for 40% of annotated CGIs in the genome. Additionally, 47% of the regions occurred at conserved regions, while only 21% overlapped both CGIs and conserved regions. The methylated regions were annotated

as being present at IGRs, PPRs, ERs, INRs, or TTRs (Appendix 4.2). Unlike the H3K4me3, the majority of DNA methylation regions were present at IGRs (19,576) and ERs (10,340) (Table 4.1). The remaining regions were distributed across INRs (8,618), PPRs (2,871) and TTRs (2,826). Further investigation of genes revealed 2,691 genes with PPR methylation, suggesting long term repression. Additionally, many genes are methylated at INRs (3,426), ERs (5,474) and TTRs (2,597). Several genes (15) lacked methylation within the gene but were methylated at both PPR and TTR.

Given the repressive role of DNA methylation, genetic variants could alter repression, resulting in changes in transcription. Therefore, like the H3K4me3 analysis, we overlapped all genetic variants to determine the underlying variation. We found the majority of the 18,768 regions with 61,554 SNVs were located at IGRs (8,438), ERs (4,149), and INRs (4,116). Additionally, 1,616 INDELS affected 987 methylated regions. While SNVs and INDELS would affect the binding of transcription factors, CNVs have the potential to duplicate or remove entire regulatory elements, thus enhancing or preventing the repressive effects. In total, 490 CNVs were found to affect 586 methylated regions, including 66 PPRs. Overall, despite potential evolutionary constraints on methylated regions, many regulatory regions are affected by a variety of types of genetic variation.

The combination of DNA methylation and H3K4me3 data allows for the analysis of differentially modified CGIs (DMC) in which CGIs are overlapping with H3K4me3 and DNA methylation (e.g., some imprinting control regions, X-inactivated genes, and genes undergoing allelic exclusion). Annotation of all 718 DMCs, as with previous steps, revealed that the majority occurred at PPRs (318), while the remainder was spread across IGRs (200), INRs (111), ERs (38), and TTRs (25) (Table 4.1). After removing the

overlapping H3K4me3 and DNA methylation peaks from the total for each modification, we found that the majority were at unique locations in the genome. The unique H3K4me3 regions were distributed across IGRs (5,872), ERs (373), INRs (3,060), PPRs (5,459), and TTRs (210). In total, 5,344 genes possessed unique H3K4me3 overlapping CGIs at PPRs. Alternatively, unique methylation regions were distributed across IGRs (18,470), ERs (10,165), INRs (8,287), PPRs (2,562), and TTRs (2,771).

Given that females were used in this study, we further investigated the DMCs on the X chromosome. Interestingly, 152 (21%) of the DMCs occurred on the X chromosome, accounting for 28% of the CGIs on the chromosome. In comparison, only 1.5% of the CGIs on other chromosomes possessed differential modifications, clearly demonstrating the detection of X-inactivation. Additionally, several genes known to undergo X-inactivation were identified as being affected by DMCs including *MAOA* and *FMR* (Figure 4.3). While many genes on the X chromosome undergo X-inactivation, some are able to escape inactivation. Analysis of several of these genes revealed that *UTX*, *ZFX*, *CRSP2*, *UBE1*, and *JARID1C* possess H3K4me3 at their PPRs, but lack DNA methylation (Figure 4.4). Outside of the X chromosome, DMCs are also an indicator of imprinting control regions and allelic exclusion of genes. The comparison of potentially imprinted genes in cattle and mice revealed several overlapping genes, respectively (Figure 4.5). Additionally, DMCs were identified in gene clusters such as a *BoLA* gene (chr23:27,748,581-27,806,566), and over 700 uncharacterized regions (Figure 4.6).

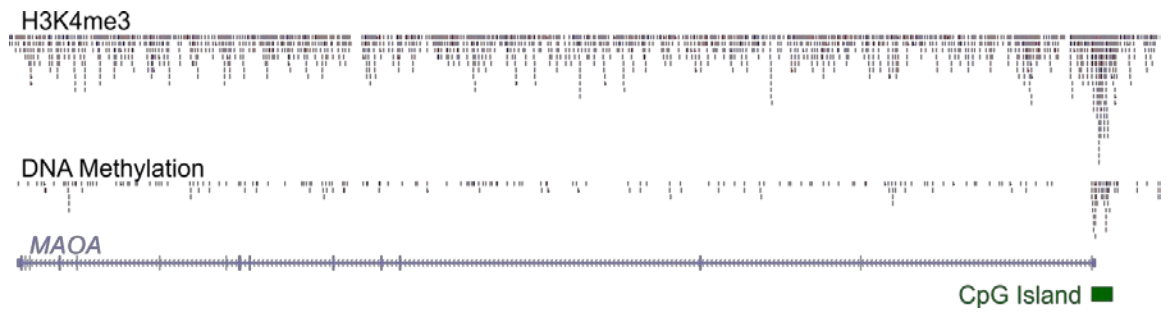


Figure 4.3 Identification of gene known to undergo X-inactivation (*MAOA*) by overlapping histone and DNA methylation

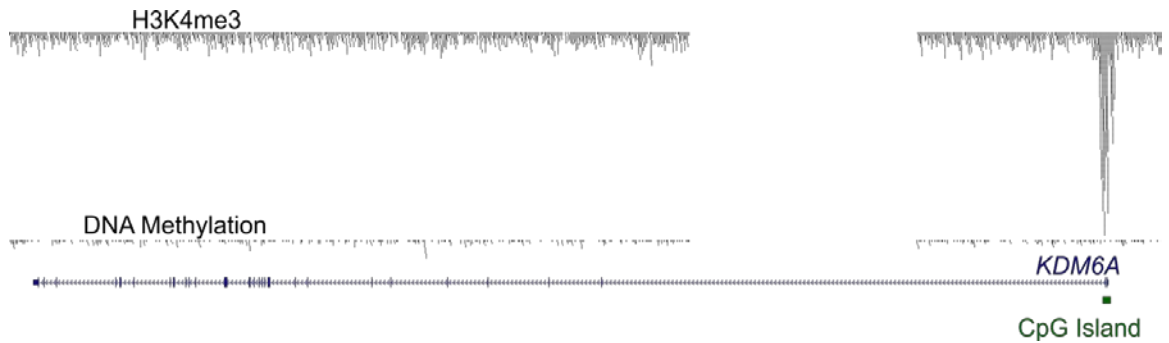


Figure 4.4 Identification of a gene known to escape X-Inactivation in humans (*KDM6A*) by histone and DNA methylation analysis in cattle

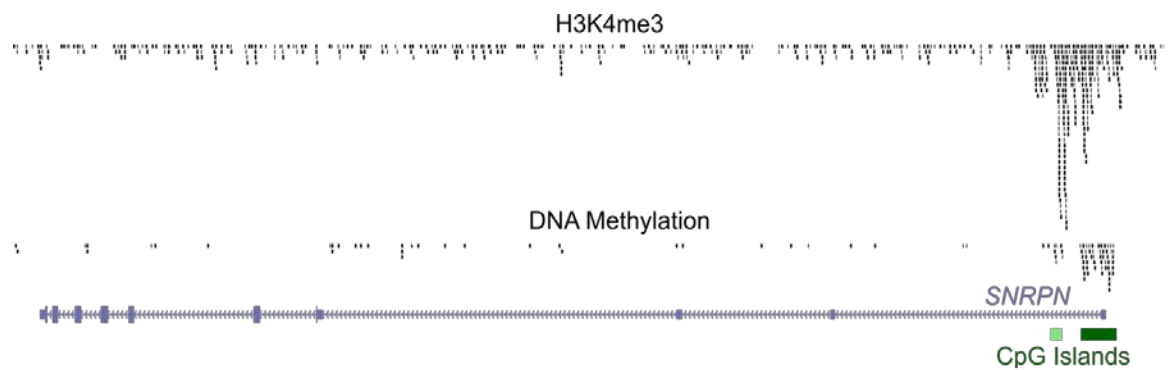


Figure 4.5 Confirmation of gene that is known to be imprinted (*SNRPN*) by overlapping histone and DNA methylation in cattle

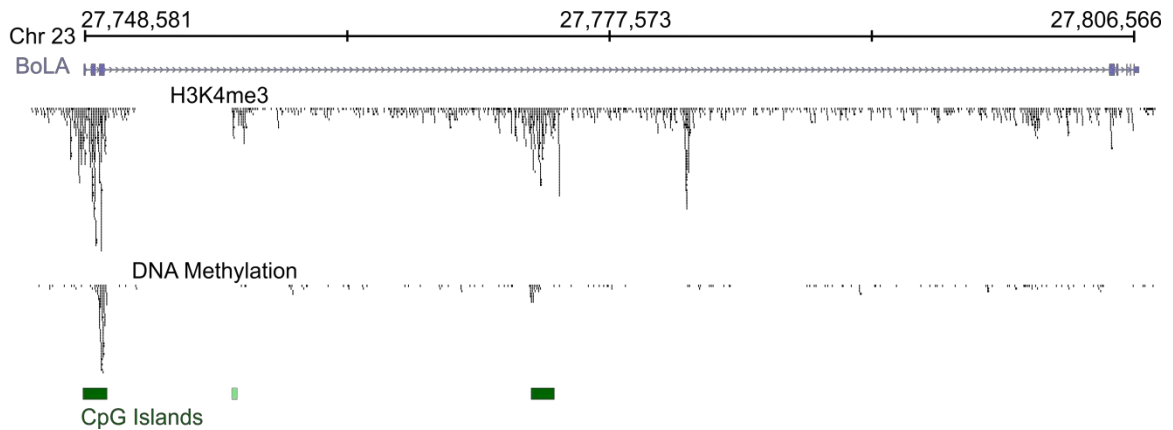


Figure 4.6 Potential allelic exclusion of *BoLA* gene indicated by overlapping histone and DNA methylation in cattle

Characterization of Variant Densities

Next, we examined the SNV densities of each genome at (IGR), intronic (INR), genic (synonymous [ER(Syn)] and nonsynonymous [ER(NS)]), PPR, RE, and DMC regions. For comparison of genetic divergence between two closely related breeds of *Bos taurus*, we determined the SNV densities of separate Angus and Holstein (*Bos taurus*) genomic sequences ([131]) (Figure 4.7, Appendix 4.3). Comparative SNV densities within each annotated region revealed that the OR of SNVs at functional coding regions (i.e., ER(NS)) and DMCs were only slightly different between the Nellore and Angus genomes relative to the Angus and Holstein genomes. As expected from the divergence of the subspecies, the highest OR of SNVs was present in IGRs. However, the SNV densities between the Nellore and Angus genomes at PPRs, REs, and TTRs were increased relative to those observed between the Angus and Holstein genomes. Collectively, these data indicate that most of the functional variation between Nellore and Angus genomes is present within regulatory elements and not at coding regions.

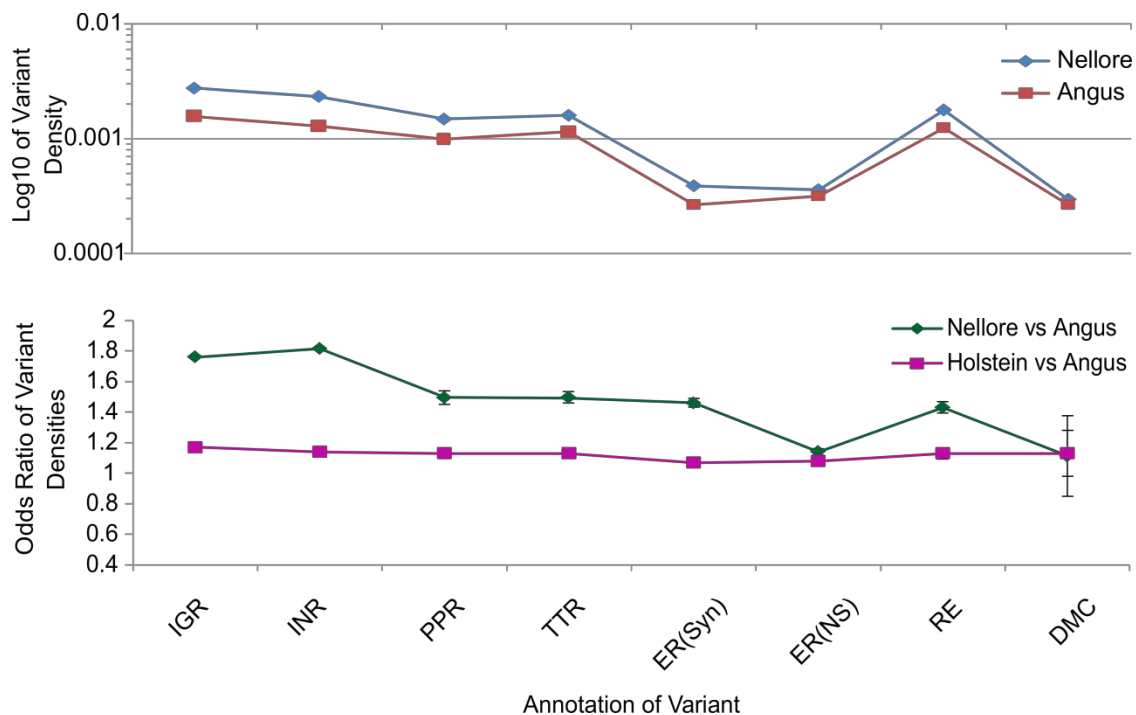


Figure 4.7 Comparison of SNVs at genetic elements. (A.) Variant densities in the Angus and Nellore genomes. (B.) Odds ratios of densities between the Angus and Nellore genomes and published Angus and Holstein genomes.

DISCUSSION

The identification of regulatory elements and their underlying genetic variation is crucial for the advancement of our understanding of complex diseases and traits in cattle. We describe the first genome-wide mapping of H3K4me3 and DNA methylation in bovine WBCs. While other histones and additional tissues would increase the numbers of elements identified, we were able to identify 6,960 intergenic regulatory elements and 5,817 actively modified PPRs. Interestingly, nearly 40% of regulatory elements fell outside of CGIs and conserved regions. This finding suggests that previous methods predicting REs based on sequence conservation and CGIs would miss a large fraction of existing elements, especially bovine-specific regions.

In addition to active modification, the repressive modification, DNA methylation, has the potential for large and long-term repression of genetic elements. These changes can lead to either enhanced or repressed gene expression. Therefore, our first analysis of more than 44,000 methylated regions in the bovine genome provided a glimpse into the repressive regulation in cattle. As seen in other species, DNA methylation was enriched for coding exons and intergenic regions due their lower repressive function and the potential role of splicing and elongation [87, 202, 203]. DNA methylation was also enriched across the X chromosome due to X-inactivation, providing the first insight into chromosomal inactivation and genes escaping the inactivation. While the functional roles of the regions will require further investigation, these data provide a starting point for better understanding of diseases and traits.

The overlap of histone and DNA methylation are linked to imprinting and allelic exclusion. DMCs often exist in highly regulated, developmentally important genes. The comparison of our histone and DNA methylation data allowed for the identification of over 700 DMCs that not only demonstrate the widespread level of differential modification in the bovine genome but provide evidence for allelic exclusion in the MHC region of cattle.

The regulatory and functional importance of promoters, intergenic regulatory elements, and DMCs also suggest a significant functional change due to any underlying genetic variation. The presence of SNVs and small INDELS within these regions could prevent or enhance the binding of factors specific for the element. These changes would in turn change the gene expression, potentially altering a trait including susceptibility to disease. Also, CNVs have the potential to cause even more dramatic expression differences due to the complete removal or duplication of regulatory

elements. Therefore, our finding that many of these regions possess SNVs, INDELS, and CNVs in the Angus and Nellore samples may help explain the lack of coding mutations, despite the large number of phenotypic differences observed in cattle.

To further understand the extent and potential role of genetic variation underlying regulatory regions, we compared the Angus and Nellore variant densities across IGRs, INRs, PPRs, TTRs, ERs, REs (intergenic H3K4me3), and DMCs. As expected, ERs and DMCs had the lowest densities, while IGR, INRs, PPRs, and REs were affected by more variants. However, ORs of the Nellore and Angus densities demonstrated that the Nellore possessed greater enrichment of SNVs in all regions except ER(NS)s and DMCs, where there was no significant difference in densities. Furthermore, the comparison of the densities of publicly available SNVs from a single Angus and Holstein cow demonstrated no significant differences in ORs at any region. Together, these data suggest that regulatory element variation may play a larger role in diversification of the subspecies than coding variation. Therefore, the focus on coding changes within QTLs and regions of selection in previous studies may have prevented the identification of true causal mutations within regulatory regions.

CHAPTER V

CNV IMPUTATION

INTRODUCTION

The recent surges in genotyping and CNV analyses using the bovine 50K and BovineHD SNP beadchips have resulted in thousands of cattle being genotyped for linkage and CNV studies. The inclusion of over 770,000 probes on the BovineHD array has led to several high-density CNV and genotyping studies in cattle [126, 204, 205]. Results of human studies have identified at least a few limitations of SNP arrays. First, SNP arrays only contain probes where a polymorphic SNP occurs in a unique sequence [24, 60, 61]. Given the repetitive nature and level of segmental duplications, these regions are underrepresented with very low SNP densities [24, 61]. Second, approximately 20% of all CNVs are not imputed by SNPs in humans and not all types of CNVs are equally imputed. Multi-allelic CNVs, where several copy numbers exist, are rarely tagged due to the recurring and complex events [24, 61]. Third, duplications tend to have lower LD with SNPs than deletion events, with rare variants having even less LD [24, 60, 61]. Despite the ability to genotype nearly 80% of CNVs, the majority of LD values are below 1 [24, 60, 61]. Taken together, these limitations suggest that a combined approach using additional methods should be used when determining individuals' genotypes in order to account for variation from both SNPs and CNVs.

Despite the recent human data demonstrating the limitations of SNP arrays, bovine genomics studies remain highly dependent on SNP arrays. The original 50K SNP array, with a limited resolution across the entire genome, has been used to genotype thousands of cattle for SNPs and CNVs. The analysis of CNVs using this

array has been reported in 6 studies looking at over 8,000 cattle [120-125]. However, these studies are only able to identify an average of two CNVs per cow. Also, the 50K SNP array can only identify CNVs down to 50 kb, with an average length of over 200 kb. The release of the BovineHD SNP beadchip has led to two large scale studies investigating CNVs at a higher resolution than the 50K array [126, 127]. The higher resolution of the BovineHD array has made it possible to detect CNVs as small as 1,018 bp in 770 cattle. Despite the wide usage of bovine SNP arrays, their major limitations have largely been ignored in cattle research. First, the design of the array allows for probes to only be placed in highly unique regions that contain a polymorphic SNP. Second, the probe spacing, like prior aCGH, lacks the resolution to detect small coding variants. Third, the many CNV genotypes cannot be imputed by SNP genotypes such as reoccurring and complex CNVs. While the extent of these limitations is still unknown in cattle, their utilization likely results in missing genotypic differences that may underlie phenotypes.

Through the comparison of BovineHD SNP genotypes and CNVs with CGH data, we demonstrated that nearly 30% of CNVs cannot be accurately imputed by the SNP array. These missed genotypes are often within exons and would be expected to cause differences in RNA expression. Furthermore, we demonstrated a combined approach that merged whole-genome sequencing, CGH, and SNP arrays that resulted in the identification of additional regions of selection and potential candidate regions for traits in cattle.

METHODS

BovineHD BeadChip CNV Analysis

The BovineHD SNP beadchip (Illumina, San Diego CA) was used to genotype all 8 Angus and Nellore samples. We received RAW data from the experiments performed by GeneSeek of Neogen Corporation (GeneSeek, Lincoln, NE). We analyzed the raw data using the Genotyping module of Illumina's GenomeStudio software (Illumina, San Diego CA). All samples were clustered based on their SNP genotypes, using Absolute Clustering. We also imported custom columns with Baylor 4.0 mapping positions for the SNPs. All genotype data were exported from GenomeStudio using the Final Report function to create a single file per sample.

Next, we identified CNVs across the bovine genome using CNVPartition 3.1.6 plugin in GenomeStudio using the following parameters: minimum probe count, 3; minimum homozygous region size, 1000000; confidence threshold, 35; include sex chromosomes, true; Detect extended homozygosity, true; and exclude intensity only, false. The CNVs identified were exported using the CNV region report function in GenomeStudio. The resulting CNVs were annotated and overlapping CNVs were merged using the join and merge functions in Galaxy (<https://main.g2.bx.psu.edu/>). All CNVs were converted from the University of Maryland Assembly 3.0 (Umd3) to Baylor 4.0 (Btau4.0) using the batch coordinate conversion (liftOver) tool on the University of California Santa Cruz's Genome browser with the following parameters: minimum ratio of bases that must remap, 0.4; allow multiple output regions, no; and If thickStart/thickEnd is not mapped, use the closest mapped base, no (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>).

PLINK Analysis

PLINK-formatted files were created from genome studio using the PLINK plugin. The files were processed into a binary file (BFile) using PLINK (Appendix 5.1, [206], <http://pngu.mgh.harvard.edu/purcell/plink/>). The `-het` option was used in PLINK to determine the inbreeding coefficients of each sample. Furthermore, the `-homozyg` option with default settings was used to determine the runs of homozygosity (ROH) within each sample. The ROHs were compared between samples to identify common regions within each breed using the `--homozyg-group` option in PLINK.

CNV Imputation by SNPs

We analyzed CNV imputation using SNPs within identified CNVs, as well as using the nearest flanking SNPs. We extracted all CNVs in the 8 Angus and Nellore samples with the overlaid SNP genotypes. We excluded all heterozygous SNPs from this analysis due to the problems with determining which chromosome contains the SNP. The SNP genotypes of samples with CNVs were compared against the Angus reference sample. A combination of SNPs within and flanking CNVs were used to fully impute a CNV. First, the ability of a SNP to tag (predictive SNP within CNV) a CNV was determined by overlapping all SNPs from all samples with the CNVs (Figure 5.1). Only CNVs containing homozygous SNPs were able to be investigated. Of the CNVs containing homozygous SNPs, those where the reference Angus and the sample have the same genotype were considered non-tagged. To expand on the ability to tag CNVs, we investigated the genotypes of the closest flanking SNPs (Figure 5.2). Without using a distance filter, we were able to assign SNP genotypes to each side for all CNVs from the 8 samples. Any CNVs where both flanking genotypes were shared between the reference and the samples were classified as non-imputed. As a final method of CNV

imputation by SNP genotypes, the data from SNPs within and flanking the CNVs were combined, allowing for a CNV to be imputed if at least one of the SNPs were different between the reference and the sample (Figure 5.3). Finally, while a CNV may appear to be imputed in one sample, it may in fact be non-imputed in another sample. In order to correct for multiple samples, any groups of CNVs or CNVRs where at least one sample was non-imputed were collectively classified as non-imputed.

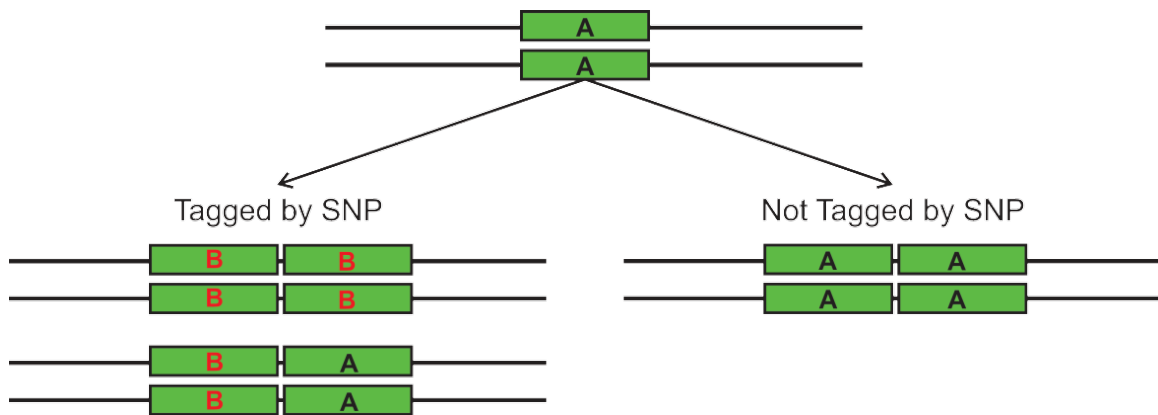


Figure 5.1 Diagram of method used to identify CNVs tagged by homozygous SNPs within the variant

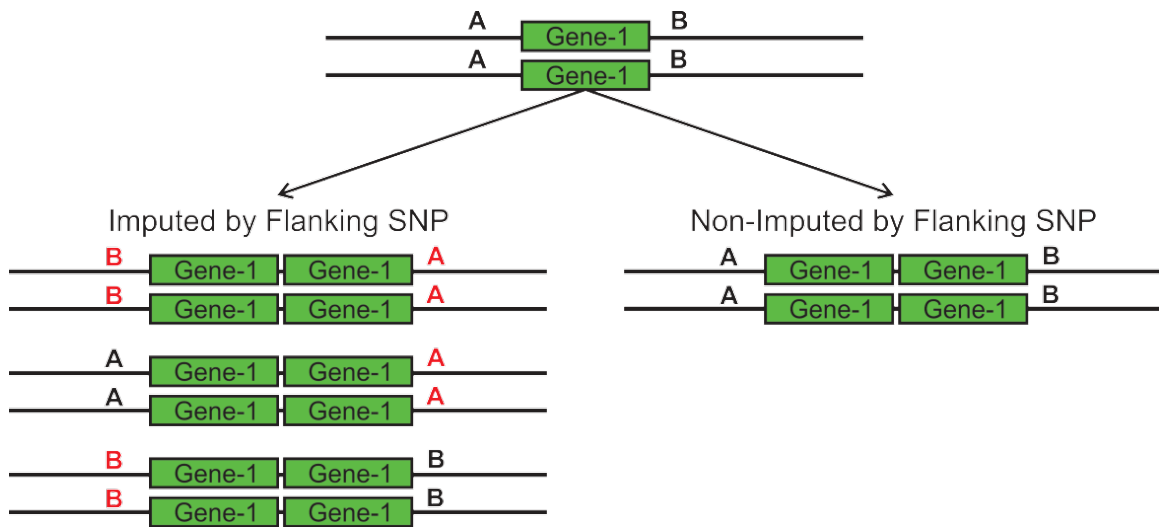


Figure 5.2 Diagram of method used to impute CNVs by homozygous SNPs flanking the variants

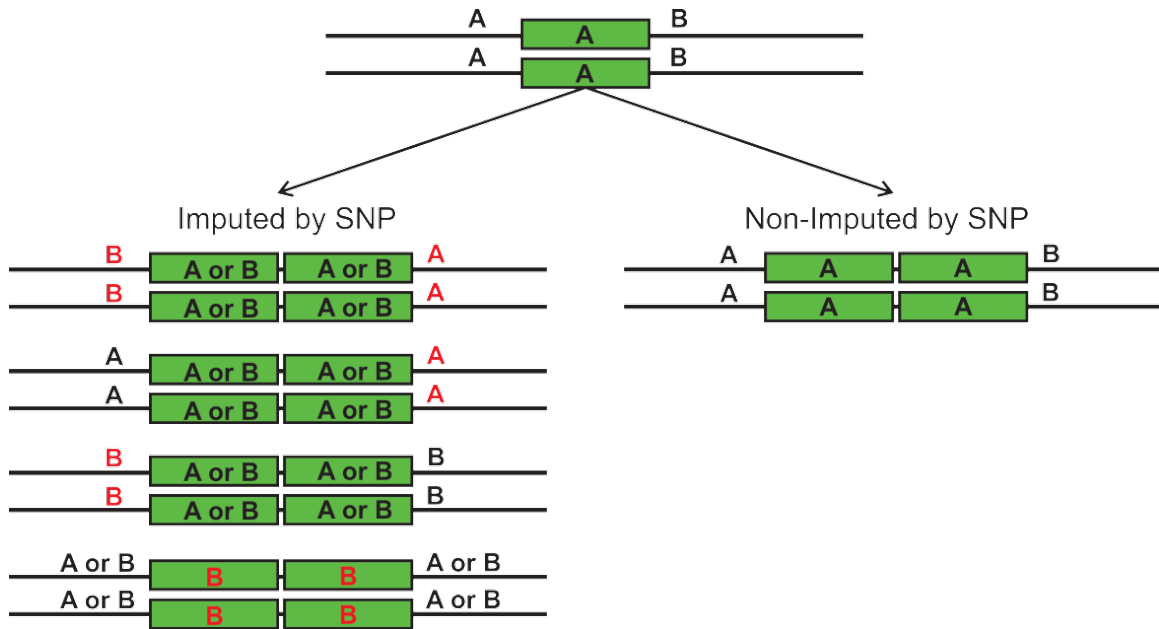


Figure 5.3 Diagram of complete CNV imputation through the combination of SNPs within and flanking CNVs

RESULTS

Genotype Analyses

The genotypic data from the BovineHD SNP beadchip were compared across all 4 Angus and 4 Nellore cattle. On average, approximately 1,000 more probes failed in the Nellore than the Angus samples. However, more than 770,000 probes yielded genotypic data in all samples. The comparison of genotypes between the breeds revealed that while more than 450,000 genotypes were shared in the Angus samples, less than 300,000 were shared between the Angus and Nellore samples (Table 5.1). Additionally, the Nellore possessed more homozygous BB alleles (~60,000) than the Angus samples. Furthermore, hierarchical cluster analysis of the genotypes demonstrated the array's ability to cluster the samples into their respective breeds (Figure 5.4).

Table 5.1 SNP genotype distributions from BovineHD array

	AA	AB	BB	Total	Comparison with Angus-4			
					Unique AA	Unique AB	Unique BB	Total Shared
Angus-4	259,833	227,149	288,257	775,239	-	-	-	-
Angus-2	263,900	219,857	291,252	775,009	90,032	126,094	93,917	466,178
Angus-3	267,047	213,730	294,577	775,354	92,528	123,688	97,038	463,194
Angus-1	270,074	207,912	297,351	775,337	90,160	118,167	93,965	474,144
Nellore-2	259,829	149,442	364,763	774,034	158,148	104,045	213,230	299,595
Nellore-3	259,149	151,255	363,722	774,126	157,582	106,137	212,444	298,923
Nellore-4	260,839	147,207	365,929	773,975	158,494	102,567	213,598	300,278
Nellore-1	258,801	151,109	363,168	773,078	156,576	106,682	211,453	299,223

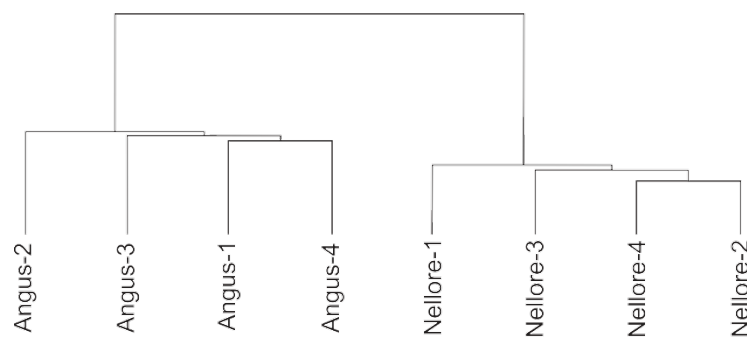


Figure 5.4 Hierarchical clustering of cattle using BovineHD SNP genotypes

To further analyze the SNP genotypes, the inbreeding coefficient was determined for each sample using PLINK. The Angus samples had a low average F value of 0.10 while the Nellore cattle appeared to be more inbred with an F of 0.39. The level of homozygosity in the samples was further investigated based on the presence of runs of homozygosity. In general, the Nellores' genomes contained nearly double the level of homozygosity (3 Mb versus 6 Mb) despite a lower number of regions per samples in the Nellore cattle (70 versus 100) (Appendix 5.2). The majority of the 697 ROH were shared with at least 1 sample, with 5 and 10 being found only in all Nellore or Angus samples, respectively

CNV Analysis

The CNVPartition script was used in GenomeStudio in order to identify 991 CNVs (608 CNVRs) in the 8 samples, 450 (328 CNVRs) of which were intergenic. The comparison the CNVs with those from the exome array highlighted the lower resolution of the SNP array, with median and minimum lengths of 21 kb and 2 kb. Overall, the CNVs were found to affect 795 ensembl genes. Unlike the results from CGH or

sequencing, functional analysis of the CNVs indicated enrichment for sensory perception and signal transduction processes.

CNV Imputation by SNP Genotyping Array

Given the extent of copy number variation throughout the bovine genome and the increasing usage of SNP genotyping arrays in bovine association studies, we investigated correlations between SNP genotypes and CNVs. SNP positions were converted to Btau4.0 coordinates and overlaid with CNVs identified by the exome array. We found that only 42% (584 of 1,376) of the CNVs identified in the Angus and Nellore cattle (n=8) had SNPs located within the boundaries of the CNVs. Using homozygous SNPs located within the CNV region, we found that the Angus had a greater percentage of missed genotypes (43% to 71%) than the Nellore (33% to 42%). The incorrect imputation of CNVRs by flanking homozygous SNPs occurred for 34% of the regions. However, the inclusion of homozygous SNPs flanking CNVs with tagged CNVs decreased the number of non-imputed CNVs to 29% (19% of CNVRs). As seen in human studies, complex regions and duplication CNVs have the lowest percentages of imputed CNVs. However, nearly 45% homozygous deletions were unable to be imputed by SNP genotypes. Therefore, while the 71% of CNVs may be predicted by SNPs, they will contain a bias toward heterozygous deletion variations.

Further analysis of non-imputed regions found that the majority lie within segmental duplications (61%) (Table 5.2). Also, while the ability of a CNV to be imputed was not found to be correlated with the length of the variant, there is a correlation when an overlap of SDs is taken into account. Of the non-imputed CNVs, 95% of those over 10 kb were found to overlap SDs (Table 5.2). Overall, 834 and 256 genes were affected by imputed and non-imputed CNVs, respectively. Functional analysis of biological

processes found enrichment for processes involved with signal transduction, sensory perception, and immunity and defense for imputed genes, while non-imputed genes were enriched for signal transduction and sensory perception (Appendix 5.3). Of the non-imputed genes, several were under selection in cattle such as *PSMB7*, *CATHL1*, *CATHL4*, *FANCC*, and *IGLL1*, further demonstrating the importance of these missed genotypes (Figure 5.5).

Table 5.2 Complete CNV imputation using genotypic data from the BovineHD SNP array

Lengths	CNVs		Imputed (%)		Non-Imputed (%)	
	Total	Overlap SDs	Total	Overlap SDs	Total	Overlap SDs
Total	1,387	677 (48.8%)	980 (70.7%)	428 (43.7%)	407 (29.3%)	249 (61.2%)
< 10 Kb	994	353 (30.5%)	703 (70.7%)	214 (30.5%)	291 (29.3%)	139 (47.8%)
> 10 Kb	393	324 (82.4%)	277 (70.5%)	214 (77.3%)	116 (29.5%)	110 (94.8%)

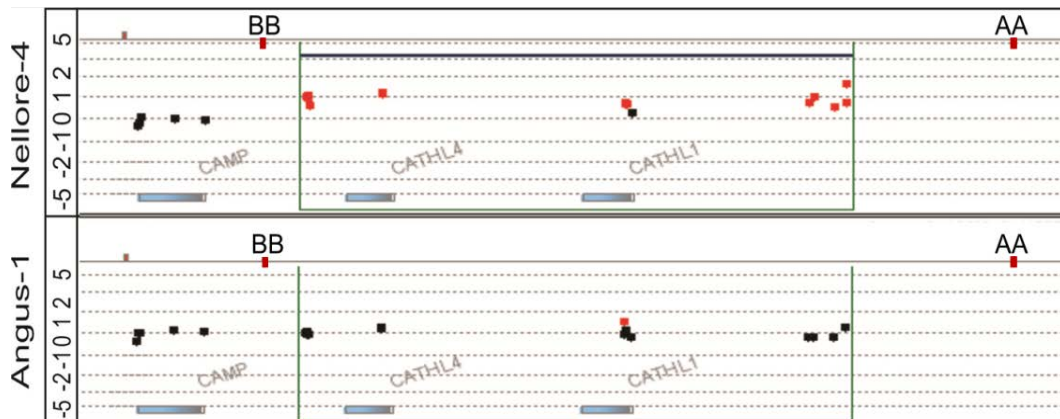


Figure 5.5 Example of CNV that is not accurately imputed by BovineHD SNP array genotypes

To further determine whether the ability to impute CNVs based on SNP genotypes could be improved through the inclusion of additional or more informative SNPs, we compared the SNVs identified by sequencing to the CNVs identified in Nellore-1 by exome aCGH. The combination of homozygous SNVs within and flanking the 321 CNVs allowed for 74% to be accurately imputed.

DISCUSSION

The recent advances in bovine genomics have largely relied on the release of the 50K SNP beadchip and more recently, the 770K BovineHD SNP beadchip. Thousands of samples and millions of dollars were spent on these designs, based on the premise the arrays capture the majority of variation in the genome. To investigate the applicability of relying solely on SNP arrays for genotypic analysis and CNV identification we compared SNP data from 8 cattle (4 Angus and 4 Nellore) with CNV data from a high-density exome array.

Based on the SNP data, the Nellore were much more inbred with 3 times more homozygous SNPs. Additionally, the Nellore cattle were found to contain fewer, but

larger, regions of homozygosity, resulting in a greater overall fraction of the genome. Together, these data correspond with known ancestral bottle-necks and inbreeding in the Nellore breed. Furthermore, CNV analysis using the SNP array revealed nearly 1,000 CNVs in the samples, with more in the Nellore than the Angus. However, when compared to the exome CGH data, we found that the resolution of the SNP array prevents the identification of CNVs less than 2,000 bases in length. Therefore, given that the median size of CNVs is less than 1 kb, a larger number of CNVs are undetectable by the SNP array. Additionally, CNVs within tandem repeats and other repetitive regions were often missed due to the lack of probe coverage in these regions.

The analysis of CNV imputation using homozygous flanking and tagging SNPs demonstrated that nearly 30% of CNVs identified by a dense exome array are not accurately imputed by the BovineHD genotypes. Regions with complex variants and duplications were less likely to be imputed than regions with heterozygous deletions. Therefore, the probe placement is not the only bias introduced by SNP arrays, differences in the ability to identify specific types of CNVs will create a preference toward deletions. Despite the elimination of probe bias through SNV detection by whole-genome sequencing, many CNVs were still unable to be accurately imputed by SNV genotypes. Given the small number of samples tested, it is very likely that even more of the potentially imputed CNVs will be shown not to be imputed in future studies. Therefore, the reliance of bovine genomics studies on SNP arrays appears to risk not detecting existing genetic variation. However, these limitations could easily be overcome through the incorporation of whole-genome sequencing and high-density CGH analyses.

CHAPTER VI

CATHELICIDIN ANALYSIS

INTRODUCTION

Copy number variants have been linked to a variety of traits and diseases in human and domestic animals. Duplications and deletions within immune related gene regions (e.g., defensin clusters and the MHC) have been linked to a variety of phenotypes and disease in several species [207-212]. CNVs cause at least 18% of the expression differences observed between individuals [213]. Therefore, duplications of immune related genes are likely to result in an increase of gene expression. If additional regulation controls are not present in the affected pathway, an increase of expression can result in a biological change. However, expression of genes producing antimicrobial peptides affects the immune traits directly, without further downstream steps in the functional pathway.

The human cathelicidin gene (*CAMP*) possesses broad-spectrum anti-microbial properties against bacteria, viruses (e.g., vaccinia, HSV), and yeast (e.g., *Candida albicans*) [214]. The expression of cathelicidin increases in response to external stimuli such as microbial and physical stresses [215]. The mRNA is translated into an inactive peptide and proteolytically cleaved into a small active peptide consisting of 37 amino acids and stored in lamellar bodies of keratinocytes and neutrophils [214, 215]. The antimicrobial properties *CAMP* led to numerous studies linking the active peptide to disorders affecting the skin and digestive tract, as well as general immunity [216]. Cathelicidin gene expression is predominately controlled by the combination of an Alu SINE and a vitamin D response element (VDRE) in the 5' UTR region of *CAMP* [217, 218]. The existence of the VDRE region allows for the induction of expression in

response to an increase in the active form of vitamin D3 (1,25(OH)2D3) [217, 219]. The induction of *CAMP* through this pathway is well documented in humans and can be achieved using a variety of stimuli.

Recently, studies in humans and mice have revealed two more distinct and independent pathways that can regulate the expression of cathelicidin without utilizing the VDRE. The farnesoid receptor (FXR) is a ligand activated transcription factor that regulates the expression of many genes in response to concentrations of bile acids [220]. The farnesoid receptor exists in many species with varying levels of conservation leading to differences in induction by several bile acids. The presence of a FXR element in the 5' region of the *CAMP* gene in humans allows for the induction of cathelicidin using a variety of different bile acids including chenodeoxycholic acid, lithocholic acid, and ursodeoxycholic acid [220, 221]. Additionally, activation of cathelicidin expression with a combination of 1,25(OH)2D3 and bile acids results in a greater induction than with the individual stimuli [220].

In addition to the FXR element, endoplasmic reticulum (ER) stress induces cathelicidin expression via the NF- κ B-C/EBP α pathway [219, 222]. The C/EBP transcription factors are a family of leucine zippers that regulate cell growth and differentiation in response to ER stress. The presence of the C/EBP transcription factor in the *CAMP* UTR region allows for induction of expression through a VDR-independent pathway. Treatment of human and mouse cells and tissues with thapsigargin and tunicamycin result in an increase of cathelicidin expression [219, 222, 223]. While much less is known about the mechanism of *CAMP* induction via the C/EBP, its existence leads to the question of other currently unknown inducers existing in humans and other species.

In cattle, the cathelicidin gene is an expanded gene family with at least 11 different genes. While the function of each gene is unknown, it is predicted that the genes possess antimicrobial properties for different groups of pathogens or work collectively toward increasing the level of immune defense in cattle. There are a few preliminary studies that link the active cathelicidin peptides to antimicrobial properties, but the regulation and expression patterns are still unknown. In a recent study using monocyte cell lines from cattle, the VDRE pathway did not result in the induction of *CATHL4*, *CATHL5*, and *CATHL6* [224, 225]. The lack of induction by 1,25(OH)₂D₃ is due to cattle missing the VDRE and Alu SINE in the UTR regions of these genes [224, 225]. However, it is not clear if the lack of induction is tissue specific, if other stimuli induce the pathway, or if the other cathelicidin genes (*e.g.*, *CAMP*, *CATHL1*, *CATHL2*, or *CATHL3*) utilize the VDR pathway. Additionally, *CATHL4* is expressed in blood, lung, trachea, liver and lymph node tissues [226, 227]. While the small numbers of existing studies of bovine *CATHL4* demonstrate expression in a few tissues, the effect of duplications on cathelicidin gene expression has yet to be determined.

METHODS

Population Analysis of the Cathelicidin Duplication

The population structure of the *CATHL4* duplication was determined by investigating 52 *Bos taurus* and 40 *Bos indicus* cattle. The cattle were determined to be *Bos indicus* based on physical characteristics, not on known pedigrees. Therefore, it is possible that the cattle were crossed with *Bos taurus*. DNA was isolated from white blood cells, as previously described (Appendix 2.1). Custom Taqman primers were created for *CATHL4* and a control gene, *TFRC*, by selecting a region free of SNVs and INDELS. The genomic sequences were then repeat masked and imported for probe

selection in Applied Biosystems' (ABI) Custom Copy Number Assay Tool. The primers for both genes were tagged with FAM and manufactured by ABI. The TaqMan assays were performed in 10 μ L reactions consisting of 5 μ L of 2X TaqMan Genotyping Master Mix (ABI), 0.5 μ L of TaqMan Copy Number Assay, 20X working, 1 μ L genomic DNA (10 ng), and 3.5 μ L H₂O. The qPCR reactions were performed using the recommended settings and copy number changes were calculated, as previously described [143].

Table 6.1 Taqman PCR primers for genomic analysis of CNV

Gene	Forward Primer	Reverse Primer	Reporter	Reporter Dye
<i>CATHL4</i>	AGAAGCTTGTGGCCTCCTTTT	GACAGCTCTTCTCCATCAACCT	CCATTTCCAGGGTAGGATGACAC	FAM
<i>TFRC</i>	CTGAATAGGTTTCATTTCCCTC ACAAACC	GCCGGTCAGCTTGTGATTAAA CTTA	CTACGAGATGTATAATGACGAAA TAC	FAM

RNA Isolation

Blood was collected from 5 Angus (Angus-2, Angus-3, Angus-4, Angus-17, Angus-18) and 4 Nellore (Nellore-1 to 4) and immediately lysed using RBC lysis solution (as used in previous sections) to collect white blood cells. The WBC from 5mL aliquots of blood were flash frozen using dry ice. Total RNA was isolated from frozen WBCs using the Ambion RNA mini extraction kit (Invitrogen). In order to ensure complete removal of DNA from the samples, both on-column and off-column DNase treatments were performed. DNA contamination was removed using two incubations of 30 minutes each at 37°C with the TURBO DNA-free kit (Invitrogen).

Tissues from two Angus steers were collected during the slaughter process and flash frozen in liquid nitrogen. Samples were obtained from the following organs Liver, bone marrow, lung, heart, esophagus, skin, lymph node, and tongue. All tissues,

except for the lymph node, were from a single Angus steer. RNA was isolated from 50 mg of each tissue using the recommended Trizol method (Invitrogen). First, 50 mg of tissue was homogenized in 1 mL of Trizol. The samples were homogenized for 5 minutes at room temperature, followed by a chloroform extraction. The RNA was purified using isopropanol and ethanol washes. The RNA was resuspended in RNase free H₂O. DNA contamination was removed using the same off-column DNase treatment as in the WBCs.

Expression Analysis

RNA from WBCs and tissues were converted to cDNA using the High Capacity RNA-cDNA Kit by ABI. Negative real-time reactions (-RT) were performed using RNA, except the enzyme was replaced with RNase free H₂O. All reactions were diluted to 40 ng/μl based on the starting RNA concentrations. The ability to amplify RNA from cathelicidin genes (*CATHL1* and *CATHL4*) was tested by standard PCR, in conjunction with a control gene (*TFRC*) (Table 6.2). The ability to amplify products from the cDNA, but not in the -RT reactions, confirmed the lack of DNA contamination.

Table 6.2 Primers for expression analysis

Gene Name	Forward Primer	Reverse Primer	cDNA Product Size	Location
<i>TFRC</i>	ctgggaacaggtgaccctta	ttcccaaatacaggacag	169 bp	chr1:71,817,801-71,819,618
<i>CATHL1</i>	cgagcagtgactcaagg	ccatggctgcttgtaatcc	133 bp	chr22:52,820,035-52,820,900
<i>CATHL4</i>	aatgaagatctgggcactcg	gtgactgtccccacacactg	136 bp	chr22:52,813,814-52,814,087

The expression profiles of all tissues and WBCs were determined through SYBR green qPCR of cDNA. The concentration of 40 ng of cDNA was tested by comparing sample dilutions ranging from 8 ng to 80 ng. The average fold changes of *CATHL1* and *CATHL4* were determined through the comparison to the reference gene, *TFRC*, using the $\Delta\Delta CT$ method. Quantitative PCR used the Luminaris Color HiGreen High ROX Master mix with the recommended protocol (Table 6.3) (Thermo-Fisher).

Table 6.3 Reaction mixture for expression analysis

Reagent	Volume per Reaction
Forward Primer (10 μ M)	0.3 μ L
Reverse Primer (10 μ M)	0.3 μ L
Water	3.15 μ L
Luminaris Master Mix (2X)	5 μ L
cDNA (40 ng/ μ L)	1 μ L
Yellow Buffer (40X)	0.25 μ L
Total Reaction Volume	10 μ L

Induction of Cathelicidin Expression in Fibroblasts

Fibroblast cultures were created from ear notches of 2 cattle, an Angus-Holstein cross (#668) and Hereford (#669), using a modified fibroblast protocol [228]. The ear notches were recovered post-slaughter and shipped in sterile saline. Any large contaminants were removed by washing the samples in 1X PBS for 2 minutes with agitation. The samples were treated with 70% ethanol while all hair was removed from the skin using sterile scalpels. The cleaned skin samples were soaked in 10% Providone-Iodine for 1 minute followed by four rinses in HEPES-Buffered saline [228]. The epidermis of the skin was removed by soaking 4- to 6-mm-wide slices of skin in

0.25% trypsin for 2 hours at 37°C. The epidermis was removed from the skin when it became an opaque film. The dermal layer of skin was cut into groups of 8- 2 x 2mm squares using a sharp scalpel and placed into 6-well culture plates. A sterile glass coverslip was placed on the arranged dermal tissues. Drops of cold Dulbecco's Modified Eagle Medium (DMEM) media were placed under the coverslip. The samples were submerged in 2 mL of complete medium and grown for 7 days at 37°C with 5% CO₂ [228]. Growth medium was changed every two days. On day 12, the coverslips and tissues were removed from all wells and cells were rinsed 2 times in 4°C 1X PBS. The cells were incubated in 1 mL of 0.25% Trypsin/EDTA for 5 minutes to remove confluent cells. Approximately 60,000 cells were plated with 5 mL of complete medium in 25-cm² flasks.

Inductions were set up by plating 45,000 cells from the first passage in 2 6-well plates. The next day, cells were treated with 10- and 100-nM concentrations of vitamin D3 (D3) (1,25,(OH)₂D₃), retinoic acid (RA), RA+D3, and a control using Dimethyl sulfoxide (DMSO). A 100uM D3 stock solution was created by dissolving 50 µg D3 with 1,200 µL DMSO. A total of 3 µl of 100 µM D3 and RA was added to 3 mL of media for the 100 nM treatments. The 10 nM treatments used 0.3 µL of the stock D3 and RA with 2.7 µL of DMSO. After 2 days, cells were harvested for RNA isolation using the Ambion RNA Mini kit. Expression of *CATHL4* was characterized using the previously mentioned Luminaris qPCR method.

RESULTS

Population Analysis of Cathelicidin CNV

The duplication of *CATHL4* was found to be present in several copy number states in both *Bos taurus* and *Bos indicus* cattle (Figure 6.1A). While copy numbers in

both subspecies ranged from 2 to 5, there was an increased prevalence of higher copy numbers in the *Bos indicus* cattle. The normalized fold changes were able to be grouped using visual inspection and standard deviations. Overall, the copy numbers were grouped into 4 copy number states (Figure 6.1B). The median copy number groups in the *Bos taurus* and *Bos indicus* were 2.5 and 4, respectively.

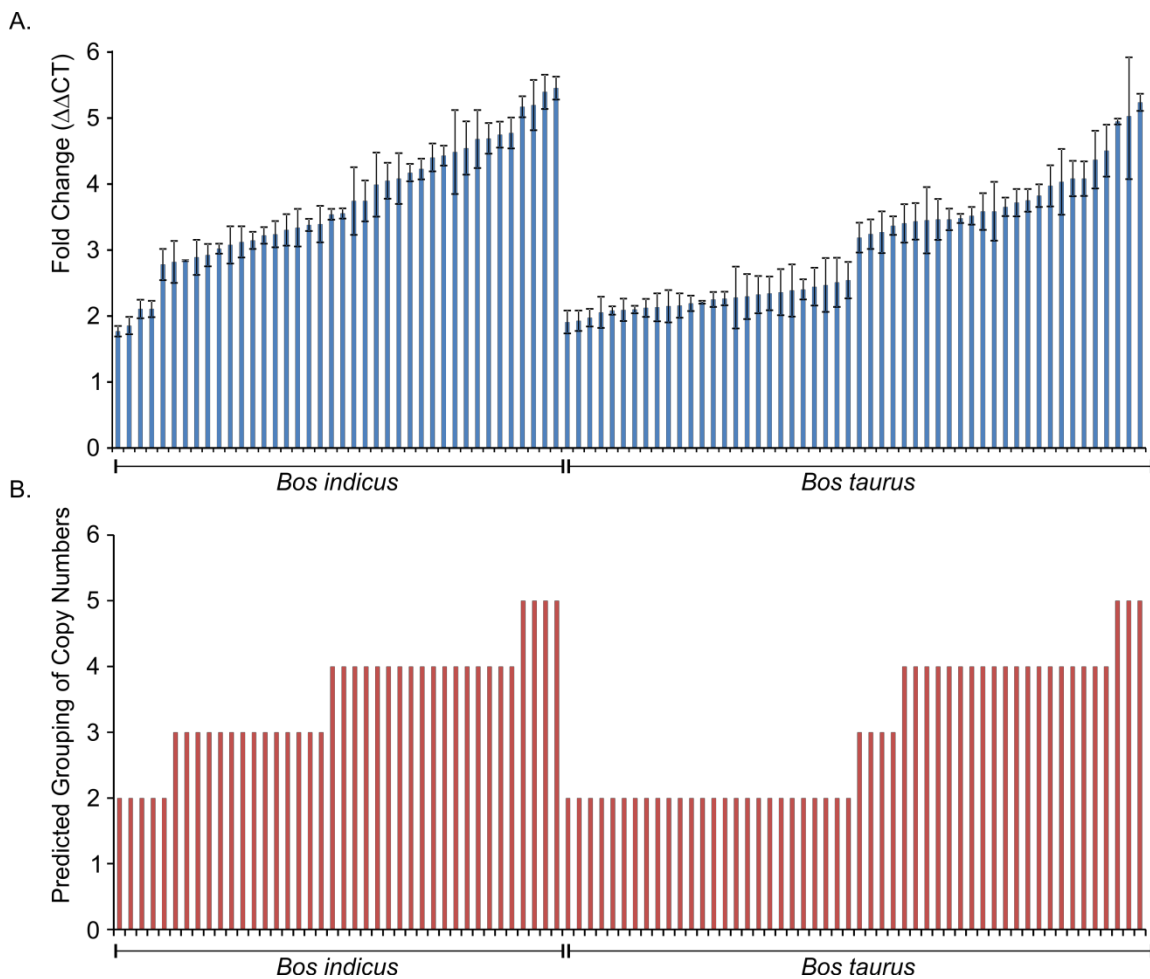


Figure 6.1 Population analysis of cathelicidin CNV in *Bos indicus* and *Bos taurus* cattle. (A.) Normalized fold changes *CATHL4* in all samples. (B.) Predicted clustering of samples into copy number groups.

Expression Profile of *CATHL1* and *CATHL4*

The expressions of *CATHL1* and *CATHL4*, both located within the cathelidicin duplication, were tested in lung, liver, tongue, esophagus, heart, skin, bone marrow, and lymph node tissues. The testing of PCR primers suggested that the expression among many of the tissues was at a basal level. Therefore, a series of cDNA dilutions from 8 ng to 80 ng per qPCR reaction demonstrated the accuracy of 40 ng concentrations. Additionally, CT values of the dilutions were much higher than the control gene, ranging from 28 to 35; however, the fold changes among the dilutions were consistent. The quality of the cDNA was confirmed by the ability to amplify the control gene in all samples. The *CATHL1* gene was expressed at high levels in the lung tissues, while at basal levels in the esophagus, heart, liver, and tongue (Figure 6.2 A). The skin was not expressing *CATHL1*. The *CATHL4* gene was highly expressed in bone marrow, with lower but elevated expression in lung, liver, and lymph node tissues. The esophagus, heart, skin, and tongue basally expressed *CATHL4* (Figure 6.2 B). The cultured fibroblasts were not found to express *CATHL4*. Also, the expression of *CATHL4* was not induced through treatments of vitamin D3 or retinoic acid.

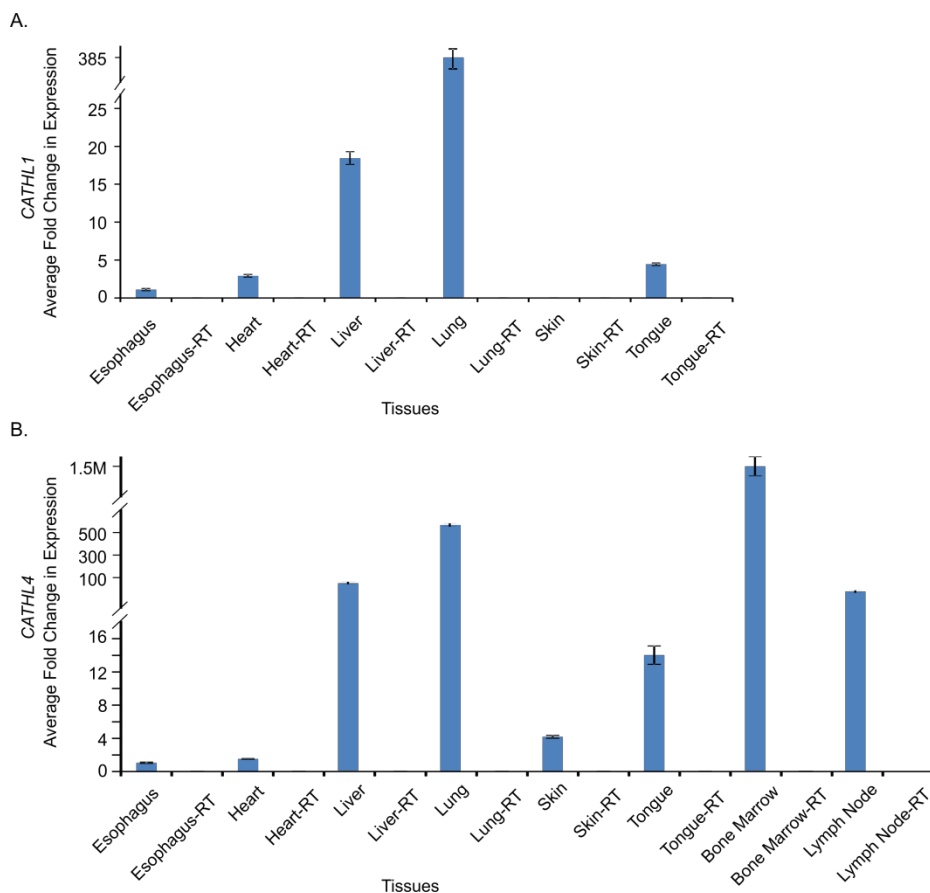


Figure 6.2 Expression profiles of (A.) *CATHL1* and (B.) *CATHL4* across several tissues from Angus cattle

Association *CATHL1* and *CATHL4* Duplication with Expression

The expression of *CATHL1* and *CATHL4* occurred at basal level in WBCs in both the animals with (Nellore) and without (Angus) the duplication. At basal levels, there was no correlation with differences in gene expression with the copy number states. Cathelicidin-1 was expressed at higher levels in Angus-3 and Angus-17, while Angus-3, Angus-17 and Nellore-2 had elevated expression of *CATHL4* (Figure 6.3). These slight differences could have been due to many factors such as the health of each animal at the time of blood collection, animals' ages, or the time of collection.

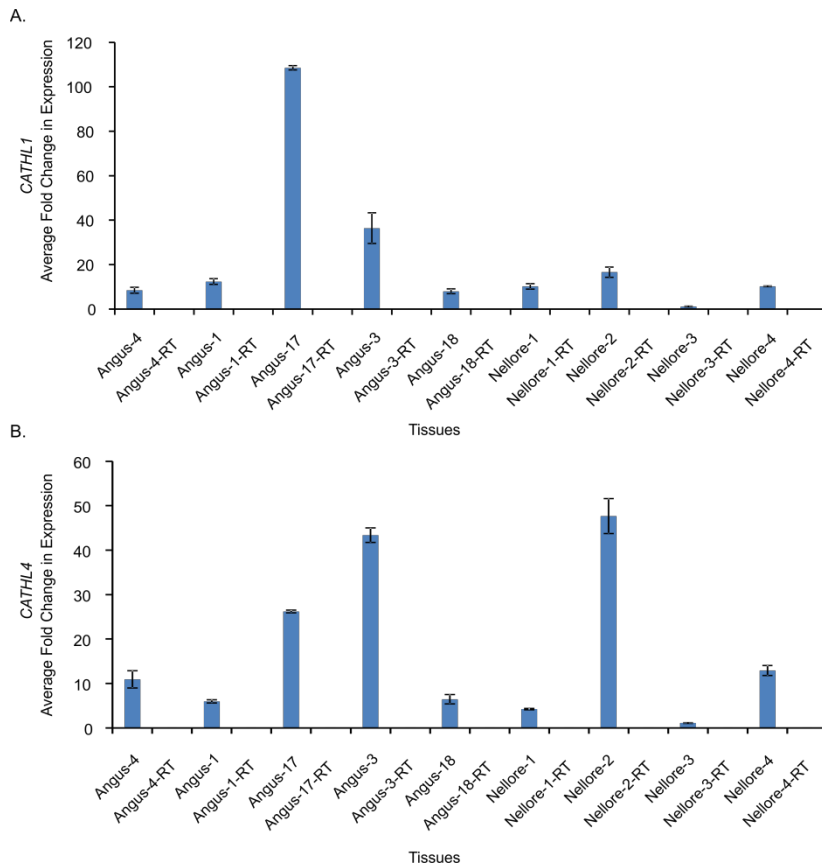


Figure 6.3 Expression profile of (A.) *CATHL1* and (B.) *CATHL4* in white blood cells of cattle with (Nellore) and without (Angus) the cathelicidin CNV

DISCUSSION

Cathelicidin is a known antimicrobial gene in humans that has undergone expansions into a gene family in cattle. The antimicrobial properties of these genes and the presence of further duplications in a portion of cattle make this region a candidate for differences in immunity observed in cattle. The limited number of cattle tested in this study suggests that *Bos indicus* have a higher frequency of gene duplications than *Bos taurus* cattle. Furthermore, the *CATHL1* and *CATHL4* genes existed in several copy

number states in both subspecies, which suggested that the CNV is recurrent in populations and was under positive selection.

The expression profiling expanded on previous studies that demonstrated the presence of gene expression by comparing relative gene expression levels between several tissues. Both genes, *CATHL1* and *CATHL4*, were basally expressed in the majority of tissues tested, which suggested that expression may need to be induced by a stressor. However, the bone marrow and lungs expressed the genes at a much higher level. The expression of cathelicidin in the lung tissue could be the result of a mild infection in the sample or a heightened defense mechanism against infection. With respiratory disease being prevalent in cattle, the expression of cathelicidin in the lungs would be an important candidate for improving resistance.

The differences in expression of genes due to the cathelicidin CNV could not be confirmed in WBCs of Angus and Nellore cattle. The genes were basally expressed in the cells and differences observed were unlikely to be a result of the CNV. Furthermore, the isolation of WBCs did not select for a single cell type, therefore, it is possible that expression is much higher in specific cell types. The resulting dilution of expression by other cell types could result in the lower level of expression.

The basal levels of expression in WBCs will need to be overcome through the use of other cell populations, such as lung or bone marrow. Alternatively, expression could be induced in cultured cells. However, we demonstrated that vitamin D3 and retinoic acid were unable to induce cathelicidin expression in fibroblasts. Therefore, it is possible that cattle possess a novel mechanism of induction. Further experiments will require testing of other compounds for their ability to induce expression in various cell lines.

Collectively, our characterization of *CATHL1* and *CATHL4* expression profiles in a several cell types revealed basal levels of expression in most samples. Additionally, we were unable to determine a correlation between expression and the presence of the cathelicidin duplication. Finally, the inability to include expression in fibroblasts, suggests that novel regulatory mechanisms control the expression of *CATHL1* and *CATHL4* in cattle.

CHAPTER VII

CONCLUSIONS

Bos taurus and *Bos indicus* represent phenotypically diverse subspecies of cattle. The *Bos taurus* subspecies consists of the predominate breeds for both milk and meat production in the United States (*i.e.*, Angus and Holstein). The *Bos indicus* subspecies (*i.e.*, Nellore and Brahman) was selected for heat tolerance and disease resistance due to their tropical origins. In the present study, we demonstrated a high level of variation between the Angus, Holstein, Nellore, and Brahman breeds through an integrated genomic variation study using aCGH, SNP BeadChips and next-generation sequencing.

While our integrated analysis was able to minimize many of the existing limitations of previous studies, we acknowledge several limitations in our array and sequencing analyses. The CGH analyses were limited due to the majority of the cattle only being analyzed using the exome array and not on the tiling array. The exome array design focused on coding regions of the genome, thereby missing any variation lying completely within introns or intergenic regions. Also, while we were able to far exceed any previous studies in terms of CNV resolution, we were still unable to confidently identify CNVs below 223 bp in length. Also, despite our dense tiling across the majority of genes in the genome, the exome array did not cover several protein- and RNA-coding genes in the bovine genome. Finally, since the array was designed using the bosTau4.0 genome assembly, genes that were unique to the Umd3.1 assembly or missing from the assembled genome were not analyzed by the CGH arrays. Similar to the CGH array analyses, the SNP analysis was limited due to the previously mentioned limitations with the BovineHD SNP array, such as the lack of probes within coding

regions and low-density tiling. Also, as we only investigated 8 of the 28 samples on the BovineHD SNP array, we were unable to perform population analyses from the SNP genotypes.

There are several limitations of our whole-genome sequencing analyses. Because we only looked at 2 samples, we cannot determine whether our results were truly representative of the breeds or simply these specific samples. Therefore, we were unable to determine population frequencies of any variant. While this is a limitation, to date, there are no large-scale genomic sequencing studies in cattle. The analysis of SNVs was limited because our depth of coverage was below the recommended 30X coverage. Also, as seen in human studies, our exome coverage was lower than within intergenic regions, making it more difficult to identify variants within coding regions. Overall, our stringent criteria resulted in large heterozygous under-call and false negative rates. However, these issues would equally affect both of the samples in our analysis and previous studies would be expected to have similar limitations.

The identification of CNVs by sequencing read-depth had several limitations that were also likely to exist in previous studies, even if they were not addressed. First, none of the programs (MrFAST, Control-FREEC, and CNV-Seq) could identify CNVs as small as our exome CGH array. The low level of overlap between the algorithms for detecting CNVs from sequencing data and with array-based CNVs suggests limitations with the algorithms. The differences in CNVs were, in part, due to uniformity and depth of read coverage, resulting in a much higher FDR and FNR than CGH methods. Despite these limitations, we were still able to identify a large number of intergenic and genic CNVs by whole-genome sequencing.

The final set of limitations occurred in our epigenetic analysis of DNA methylation and H3K4me3 in WBCs. It is likely that many regulatory regions were missed and could be detected by the inclusion of additional histone modifications and tissue types. Also, the addition of more cattle would allow for the comparison of epigenetic profiles among cattle. Finally, the analysis of DNA methylation could be improved to a base-pair-level-resolution through the use of bisulfite sequencing.

Despite several limitations from our analyses, we were able to perform a comprehensive comparative analysis of genomic variation between *Bos taurus* and *Bos indicus* genomes. We created and utilized the first high resolution exome tiling array. The probe selection allowed for an analysis of small CNVs within exons and tandem repeats that far exceeds previous studies. The analysis clearly showed a high level of structural variation between cattle with over 700 CNVRs affecting more than 1,300 genes. In addition, we found that CNVs are highly enriched for lengths well below the resolution of previous CGH and SNP array designs. Functional analysis of genes affected by CNVs identified significant enrichment for several processes including immune and defense processes. Furthermore, many genes were found to be under selection within the breeds and subspecies, resulting in several candidate genes for the diversification of traits.

The use of SNP arrays for CNV studies is rapidly increasing due to lower costs and the potential benefits of combining SNP and CNV analyses. However, the numerous potential limitations with these analyses are largely ignored despite numerous reports of these limitations identified in human studies. Therefore, we used the BovineHD SNP array to analyze 8 samples for CNVs and the ability to impute CNVs from our CGH analyses. While the SNP array identified numerous CNVs, many of the

CNVs did not overlap with known CNVs from the CGH analyses. The poor overlap between the platforms highlights the limitations from probe placement, density, and genome assemblies. Also, the ability of SNP genotypes to impute CNVs (from exome array) was tested using the closest flanking SNPs and those within the CNVs. The combination of these methods resulted in nearly 30% of the CNVs being incorrectly imputed by the SNP array. There are many potential reasons for these discrepancies. The first possible error could be a result of the FDR in the exome array; however, given the low FDR of the array, false discoveries were unlikely to result in the large discrepancy of CNV imputation. Furthermore, CNVs, such as the cathelicidin duplication, that were confirmed using qPCR, were found to be incorrectly imputed by the SNP array. Another possible error could be linked to the ability of probes to accurately genotype regions within repetitive and segmentally duplicated regions, as was observed in human studies. We predict that SDs and probe placement are the main issues with CNV imputation, given the majority (95%) of incorrectly imputed CNVs that were > 10 kb were located within segmentally duplicated regions. In conjunction with SDs, recently duplicated regions and commonly duplicated regions affect imputation because the SNP may have arose before the CNV or the CNV occurs on several allelic backgrounds. Finally, given that we only investigated 8 samples, it is possible that many of the CNVs that we defined as being imputed, will not be imputed through the investigation of additional samples. Given these and many other possible sources of mis-imputation of CNVs by SNP arrays, it is likely that relying solely on these arrays for large-scale GWAS and linkage studies could potentially lead to problems due to missing genotypic differences from CNVs. Collectively, this study demonstrates the

need for integrated CNV approaches combined with whole-genome sequencing data to identify novel variants that remain undetected by both array methods.

Our genome-wide analysis of a single Angus and Nellore cow allowed for the expansion of our CNV analyses by identifying more than 10 million SNVs, 300,000 INDELs, and 900 CNVs across the entire genomes. The biological process enrichment of SNVs and CNVs between the *Bos indicus* and *Bos taurus* comparisons identified processes involved in immunity and defense, suggesting an important role in the diversification of immune traits. Despite the enrichment for immune related genes, the majority of genes were identical at the amino acid level between the subspecies. This suggests that while coding changes play a role in phenotypes, the low level of variation is unlikely to account for all of the differences observed between the subspecies.

To further identify variation underlying phenotypes in cattle, we investigated non-coding regulatory regions, such as promoters and intergenic regulatory elements. The combination of H3K4me3 and DNA methylation sequencing provided the first epigenetic maps of regulatory regions in WBCs of cattle. While previous studies have utilized comparative genetic approaches to identify regulatory regions, we found that nearly 40% of regulatory regions were located outside of conserved regions. Therefore, we suggest that while comparative approaches may capture conserved regulatory regions, bovine specific elements may be missed. Using an unbiased approach of epigenetic profiling, we were able to identify nearly 7,000 intergenic regulatory elements and thousands of genic regions that are actively, repressively or differentially modified. Additionally, we were able to demonstrate the first genome-wide mapping of putative imprinted and allelically excluded regions. Because the analysis was performed on a

female sample, we were also able to identify genes undergoing and escaping X-inactivation.

The annotation the non-coding regions of the genome using our epigenomic maps allowed for the comparisons of variant densities across the genome. As expected, we observed an increase in variant density in the Nellore across all regions of the genome. Also, regions known to have strong selective pressures, coding regions and differentially-modified CpG islands, have extremely low variant densities in both subspecies. A comparison of the odd ratios between the Angus and Nellore demonstrated that nsSNVs and DMCs have no significant differences in variant densities, while REs and PPRs have significant increases in variant densities in the Nellore. In order to determine if the difference in ORs represent differences between individuals or actual differences between the subspecies, SNVs from a previously published Angus and Holstein, both *Bos taurus*, were annotated and ORs were compared to the Angus vs Nellore ORs. The variant densities at all regions were nearly identical between the Angus and Holstein, with no significant differences. Collectively, these findings, combined with our dN/dS analysis, suggested that non-coding regions may be playing a greater role in phenotypic variation between subspecies than coding variants. These findings were further supported by the vast amount of data from the recently released human ENOCDE project. The ENCODE project suggested quantitative traits are usually caused by regulatory element variation while Mendelian traits are caused by coding mutations.

We placed all of our data on a public server to facilitate access by other investigators. All of the data can be visualized using an IGV viewer and any of the custom tracks can be downloaded for further analysis. Additionally, we have created a

database of known mutations that are causal and associated with traits and diseases. This database will be a valuable resource for quickly genotyping cattle using whole-genome sequencing data.

CNVs were found to affect the largest amount of genetic variation in the bovine genome with over 60 Mb being affected, while SNVs only affected 7 Mb. Additionally, we provided the first epigenetic maps of histone and DNA methylation across the bovine genome. CNVs affected both coding and regulatory regions of the bovine genome. Despite the enrichment for immune processes affected by nsSNVs, analyses of variant densities suggest that non-coding variation at regulatory elements may underlie many of the diverse traits between the two subspecies. We expect that these findings will become a valuable resource for directing future studies aimed at better identifying genetic causes underlying traits in cattle. We also provide numerous candidate regions and mutations for a variety of traits in cattle. While the aim of this study was not to identify causal mutations, future studies should investigate their functional role.

REFERENCES

1. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome*. Nat Rev Genet, 2006. **7**(2): p. 85-97.
2. Reich, D.E., et al., *Human genome sequence variation and the influence of gene history, mutation and recombination*. Nat Genet, 2002. **32**(1): p. 135-42.
3. Przeworski, M., R.R. Hudson, and A. Di Rienzo, *Adjusting the focus on human variation*. Trends Genet, 2000. **16**(7): p. 296-302.
4. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing*. Nature, 2010. **467**(7319): p. 1061-73.
5. Shen, H., et al., *Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians*. PLoS One, 2013. **8**(4): p. e59494.
6. Gonzaga-Jauregui, C., J.R. Lupski, and R.A. Gibbs, *Human genome sequencing in health and disease*. Annu Rev Med, 2012. **63**: p. 35-61.
7. Olsson, M., et al., *A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs*. PLoS Genet, 2011. **7**(3): p. e1001332.
8. Fondon, J.W., 3rd and H.R. Garner, *Molecular origins of rapid and continuous morphological evolution*. Proc Natl Acad Sci U S A, 2004. **101**(52): p. 18058-63.
9. Redon, R., et al., *Global variation in copy number in the human genome*. Nature, 2006. **444**(7118): p. 444-54.
10. Perry, G.H., et al., *Hotspots for copy number variation in chimpanzees and humans*. Proc Natl Acad Sci U S A, 2006. **103**(21): p. 8006-11.
11. Perry, G.H., et al., *Copy number variation and evolution in humans and chimpanzees*. Genome Res, 2008. **18**(11): p. 1698-710.

12. Stewart, C., et al., *A comprehensive map of mobile element insertion polymorphisms in humans*. PLoS Genet, 2011. **7**(8): p. e1002236.
13. Beck, C.R., et al., *LINE-1 elements in structural variation and disease*. Annu Rev Genomics Hum Genet, 2011. **12**: p. 187-215.
14. Huang, C.R., et al., *Mobile interspersed repeats are major structural variants in the human genome*. Cell, 2010. **141**(7): p. 1171-82.
15. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population*. Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
16. Marques-Bonet, T., et al., *A burst of segmental duplications in the genome of the African great ape ancestor*. Nature, 2009. **457**(7231): p. 877-881.
17. Gu, W., F. Zhang, and J.R. Lupski, *Mechanisms for human genomic rearrangements*. Pathogenetics, 2008. **1**(1): p. 4.
18. Liu, P., et al., *Mechanisms for recurrent and complex human genomic rearrangements*. Curr Opin Genet Dev, 2012. **22**(3): p. 211-20.
19. Darai-Ramqvist, E., et al., *Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions*. Genome Res, 2008. **18**(3): p. 370-9.
20. Lam, K.W. and A.J. Jeffreys, *Processes of de novo duplication of human alpha-globin genes*. Proc Natl Acad Sci U S A, 2007. **104**(26): p. 10950-5.
21. Flores, M., et al., *Recurrent DNA inversion rearrangements in the human genome*. Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6099-106.
22. Steinmann, K., et al., *Type 2 NF1 deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination*. Am J Hum Genet, 2007. **81**(6): p. 1201-20.
23. Reiter, L.T., et al., *Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients*. Am J Hum Genet, 1998. **62**(5): p. 1023-33.

24. Kato, M., et al., *Population-genetic nature of copy number variations in the human genome*. Hum Mol Genet, 2010. **19**(5): p. 761-73.
25. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing*. Nature, 2011. **470**(7332): p. 59-65.
26. Lieber, M.R., *The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway*. Annu Rev Biochem, 2010. **79**: p. 181-211.
27. Hastings, P.J., et al., *Mechanisms of change in gene copy number*. Nat Rev Genet, 2009. **10**(8): p. 551-64.
28. Gemayel, R., et al., *Variable tandem repeats accelerate evolution of coding and regulatory sequences*. Annual Review of Genetics, 2010. **44**(1): p. 445-477.
29. Legendre, M., et al., *Sequence-based estimation of minisatellite and microsatellite repeat variability*. Genome Res, 2007. **17**(12): p. 1787-96.
30. Hastings, P.J., G. Ira, and J.R. Lupski, *A microhomology-mediated break-induced replication model for the origin of human copy number variation*. PLoS Genet, 2009. **5**(1): p. e1000327.
31. Zhang, F., et al., *The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans*. Nat Genet, 2009. **41**(7): p. 849-53.
32. Kazazian, H.H., Jr., *Mobile elements: drivers of genome evolution*. Science, 2004. **303**(5664): p. 1626-32.
33. Deininger, P.L., et al., *Mobile elements and mammalian genome evolution*. Curr Opin Genet Dev, 2003. **13**(6): p. 651-8.
34. Zhang, F., et al., *Copy number variation in human health, disease, and evolution*. Annual Review of Genomics and Human Genetics, 2009. **10**(1): p. 451-481.

35. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.
36. Keren, H., G. Lev-Maor, and G. Ast, *Alternative splicing and evolution: diversification, exon definition and function*. Nat Rev Genet, 2010. **11**(5): p. 345-55.
37. Lupski, J.R., *Retrotransposition and structural variation in the human genome*. Cell, 2010. **141**(7): p. 1110-2.
38. Gemayel, R., et al., *Variable tandem repeats accelerate evolution of coding and regulatory sequences*. Annu Rev Genet, 2010. **44**: p. 445-77.
39. Kondrashov, A.S., *Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases*. Hum Mutat, 2003. **21**(1): p. 12-27.
40. Lupski, J.R., *Structural variation in the human genome*. N Engl J Med, 2007. **356**(11): p. 1169-71.
41. Buard, J., et al., *Influences of array size and homogeneity on minisatellite mutation*. EMBO J, 1998. **17**(12): p. 3495-502.
42. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat*. Am J Hum Genet, 1998. **62**(6): p. 1408-15.
43. van Ommen, G.J., *Frequency of new copy number variation in humans*. Nat Genet, 2005. **37**(4): p. 333-4.
44. Lupski, J.R., *Genomic rearrangements and sporadic disease*. Nat Genet, 2007. **39**(7 Suppl): p. S43-7.
45. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors*. Science, 1992. **258**(5083): p. 818-21.
46. Schena, M., et al., *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes*. Proc Natl Acad Sci U S A, 1996. **93**(20): p. 10614-9.

47. Shalon, D., S.J. Smith, and P.O. Brown, *A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization*. *Genome Res*, 1996. **6**(7): p. 639-45.
48. Geschwind, D.H., et al., *Klinefelter's syndrome as a model of anomalous cerebral laterality: Testing gene dosage in the X chromosome pseudoautosomal region using a DNA microarray*. *Developmental Genetics*, 1998. **23**(3): p. 215-229.
49. Kraus, J., et al., *High-resolution comparative hybridization to combed DNA fibers*. *Hum Genet*, 1997. **99**(3): p. 374-80.
50. Solinas-Toldo, S., et al., *Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances*. *Genes Chromosomes Cancer*, 1997. **20**(4): p. 399-407.
51. Heller, R.A., et al., *Discovery and analysis of inflammatory disease-related genes using cDNA microarrays*. *Proc Natl Acad Sci U S A*, 1997. **94**(6): p. 2150-5.
52. Pollack, J.R., et al., *Genome-wide analysis of DNA copy-number changes using cDNA microarrays*. *Nat Genet*, 1999. **23**(1): p. 41-6.
53. Celestino-Soper, P.B.S., et al., *Use of array CGH to detect exonic copy number variants throughout the genome in autism families detects a novel deletion in TMLHE*. *Human Molecular Genetics*, 2011.
54. Boone, P.M., et al., *Detection of clinically relevant exonic copy-number changes by array CGH*. *Human Mutation*, 2010. **31**(12): p. 1326-1342.
55. Curtis, C., et al., *The pitfalls of platform comparison: DNA copy number array technologies assessed*. *BMC Genomics*, 2009. **10**: p. 588.
56. Haraksingh, R.R., et al., *Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms*. *PLoS One*, 2011. **6**(11): p. e27859.
57. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping*. *Nat Rev Genet*, 2011. **12**(5): p. 363-76.

58. Gentalen, E. and M. Chee, *A novel method for determining linkage between DNA sequences: hybridization to paired probe arrays*. *Nucleic Acids Res*, 1999. **27**(6): p. 1485-91.
59. Colella, S., et al., *QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data*. *Nucleic Acids Res*, 2007. **35**(6): p. 2013-25.
60. Schrider, D.R. and M.W. Hahn, *Lower linkage disequilibrium at CNVs is due to both recurrent mutation and transposing duplications*. *Mol Biol Evol*, 2010. **27**(1): p. 103-11.
61. McCarroll, S.A., et al., *Integrated detection and population-genetic analysis of SNPs and copy number variation*. *Nat Genet*, 2008. **40**(10): p. 1166-74.
62. Rogers, Y.H. and J.C. Venter, *Genomics: massively parallel sequencing*. *Nature*, 2005. **437**(7057): p. 326-7.
63. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry*. *Nature*, 2008. **456**(7218): p. 53-9.
64. Alkan, C., et al., *Personalized copy number and segmental duplication maps using next-generation sequencing*. *Nat Genet*, 2009. **41**(10): p. 1061-7.
65. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. *Bioinformatics*, 2009. **25**(14): p. 1754-60.
66. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. *Genome Biol*, 2009. **10**(3): p. R25.
67. Chen, K., et al., *BreakDancer: an algorithm for high-resolution mapping of genomic structural variation*. *Nat Methods*, 2009. **6**(9): p. 677-81.
68. Boeva, V., et al., *Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization*. *Bioinformatics*, 2011. **27**(2): p. 268-9.

69. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data*. *Bioinformatics*, 2012. **28**(3): p. 423-5.
70. Ajay, S.S., et al., *Accurate and comprehensive sequencing of personal genomes*. *Genome Res*, 2011. **21**(9): p. 1498-505.
71. Bickhart, D.M., et al., *Copy number variation of individual cattle genomes using next-generation sequencing*. *Genome Res*, 2012. **22**(4): p. 778-90.
72. Xie, C. and M.T. Tammi, *CNV-seq, a new method to detect copy number variation using high-throughput sequencing*. *BMC Bioinformatics*, 2009. **10**: p. 80.
73. Li, Y., et al., *Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly*. *Nat Biotechnol*, 2011. **29**(8): p. 723-30.
74. Alkan, C., S. Sajjadian, and E.E. Eichler, *Limitations of next-generation genome sequence assembly*. *Nat Methods*, 2011. **8**(1): p. 61-5.
75. Black, J.C., C. Van Rechem, and J.R. Whetstone, *Histone lysine methylation dynamics: establishment, regulation, and biological impact*. *Mol Cell*, 2012. **48**(4): p. 491-507.
76. Hnilicova, J. and D. Stanek, *Where splicing joins chromatin*. *Nucleus*, 2011. **2**(3): p. 182-8.
77. Dhami, P., et al., *Complex exon-intron marking by histone modifications is not determined solely by nucleosome distribution*. *PLoS One*, 2010. **5**(8): p. e12339.
78. Liu, T., et al., *Broad chromosomal domains of histone modification patterns in *C. elegans**. *Genome Res*, 2011. **21**(2): p. 227-36.
79. Spies, N., et al., *Biased chromatin signatures around polyadenylation sites and exons*. *Mol Cell*, 2009. **36**(2): p. 245-54.

80. Riddle, N.C., et al., *Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin*. *Genome Res*, 2011. **21**(2): p. 147-63.
81. Kolasinska-Zwierz, P., et al., *Differential chromatin marking of introns and expressed exons by H3K36me3*. *Nature Genetics*, 2009. **41**(3): p. 376-381.
82. El-Osta, A. and A.P. Wolffe, *DNA methylation and histone deacetylation in the control of gene expression: basic biochemistry to human development and disease*. *Gene Expr*, 2000. **9**(1-2): p. 63-75.
83. Gopalakrishnan, S., B.O. Van Emburgh, and K.D. Robertson, *DNA methylation in development and human disease*. *Mutat Res*, 2008. **647**(1-2): p. 30-8.
84. Robertson, K.D., *DNA methylation and human disease*. *Nat Rev Genet*, 2005. **6**(8): p. 597-610.
85. Robertson, K.D. and A.P. Wolffe, *DNA methylation in health and disease*. *Nat Rev Genet*, 2000. **1**(1): p. 11-9.
86. Shames, D.S., J.D. Minna, and A.F. Gazdar, *DNA methylation in health, disease, and cancer*. *Curr Mol Med*, 2007. **7**(1): p. 85-102.
87. Zemach, A., et al., *Genome-wide evolutionary analysis of eukaryotic DNA methylation*. *Science*, 2010. **328**(5980): p. 916-9.
88. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome*. *Nat Genet*, 2007. **39**(4): p. 457-66.
89. Sorensen, A.L., et al., *Promoter DNA methylation patterns of differentiated cells are largely programmed at the progenitor stage*. *Mol Biol Cell*, 2010. **21**(12): p. 2066-77.
90. Scarano, M.I., et al., *DNA methylation 40 years later: Its role in human health and disease*. *J Cell Physiol*, 2005. **204**(1): p. 21-35.

91. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*. Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
92. Ward, L.D. and M. Kellis, *Interpreting non-coding genetic variation in complex traits and human disease*. Nat Biotechnol, 2012. **30**(11): p. 1095-106.
93. Trynka, G., et al., *Chromatin marks identify critical cell types for fine mapping complex trait variants*. Nat Genet, 2013. **45**(2): p. 124-30.
94. Henrichsen, C.N., E. Chaignat, and A. Reymond, *Copy number variants, diseases and gene expression*. Hum Mol Genet, 2009. **18**(R1): p. R1-8.
95. Salmon Hillbertz, N.H., et al., *Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs*. Nat Genet, 2007. **39**(11): p. 1318-20.
96. Sundstrom, E., et al., *Copy number expansion of the STX17 duplication in melanoma tissue from Grey horses*. BMC Genomics, 2012. **13**(1): p. 365.
97. Rosengren Pielberg, G., et al., *A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse*. Nat Genet, 2008. **40**(8): p. 1004-9.
98. United States. Foreign Agricultural Service. and United States. World Agricultural Outlook Board., *Livestock and poultry, world markets and trade, in Circular series / United States Department of Agriculture, Foreign Agricultural Service*2012, The Service: Washington, D.C. p. v.
99. Achilli, A., et al., *Mitochondrial genomes of extinct aurochs survive in domestic cattle*. Curr Biol, 2008. **18**(4): p. R157-8.
100. Elsik, C.G., et al., *The genome sequence of taurine cattle: a window to ruminant biology and evolution*. Science, 2009. **324**(5926): p. 522-8.
101. Bradley, D.G., et al., *Mitochondrial diversity and the origins of African and European cattle*. Proc Natl Acad Sci U S A, 1996. **93**(10): p. 5131-5.

102. Gibbs, R.A., et al., *Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds*. Science, 2009. **324**(5926): p. 528-32.
103. Hansen, P.J., *Physiological and cellular adaptations of zebu cattle to thermal stress*. Animal Reproduction Science, 2004. **82-83**: p. 349-60.
104. Hutchins.Jc and G.D. Brown, *Penetrance of cattle coats by radiation*. Journal of Applied Physiology, 1969. **26**(4): p. 454-&.
105. Bo, G.A., P.S. Baruselli, and M.F. Martinez, *Pattern and manipulation of follicular development in Bos indicus cattle*. Animal Reproduction Science, 2003. **78**(3-4): p. 307-326.
106. Chenoweth, P.J., *Aspects of reproduction in female Bos-indicus cattle - a review*. Australian Veterinary Journal, 1994. **71**(12): p. 422-426.
107. Crouse, J.D., et al., *Comparisons of Bos-indicus and Bos-taurus inheritance for carcass beef characteristics and meat palatability*. J Anim Sci, 1989. **67**(10): p. 2661-2668.
108. Finch, V.A., *Comparison of non-evaporative heat-transfer in different cattle breeds*. Australian Journal of Agricultural Research, 1985. **36**(3): p. 497-508.
109. United States. National Agricultural Statistics Service., *Cattle and calves death loss2011*, Washington, D.C.: The Service. v.
110. Kossaibati, M.A. and R.J. Esslemont, *The costs of production diseases in dairy herds in England*. Vet J, 1997. **154**(1): p. 41-51.
111. Mathews, K. *U.S. Cattle and beef industry, 2002-2011*. 2012; Available from: <http://www.ers.usda.gov/topics/animal-products/cattle-beef/statistics-information.aspx>.
112. Oliver, S.P., S.E. Murinda, and B.M. Jayarao, *Impact of Antibiotic Use in Adult Dairy Cows on Antimicrobial Resistance of Veterinary and Human Pathogens: A Comprehensive Review*. Foodborne Pathogens and Disease, 2011. **8**(3): p. 337-355.

113. Georges, M., et al., *Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing*. Genetics, 1995. **139**(2): p. 907-20.
114. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: potential next steps on a genetic journey*. Hum Mol Genet, 2008. **17**(R2): p. R156-65.
115. Liu, G.E., et al., *Initial analysis of copy number variations in cattle selected for resistance or susceptibility to intestinal nematodes*. Mamm Genome, 2011. **22**(1-2): p. 111-21.
116. Liu, G.E., et al., *Analysis of copy number variations among diverse cattle breeds*. Genome Res, 2010. **20**(5): p. 693-703.
117. Liu, G.E., et al., *Detection of germline and somatic copy number variations in cattle*. Dev Biol (Basel), 2008. **132**: p. 231-7.
118. Kijas, J.W., et al., *Analysis of copy number variants in the cattle genome*. Gene, 2011. **482**(1-2): p. 73-7.
119. Fadista, J., et al., *Copy number variation in the bovine genome*. BMC Genomics, 2010. **11**: p. 284.
120. Bae, J.S., et al., *Identification of copy number variations and common deletion polymorphisms in cattle*. BMC Genomics, 2010. **11**: p. 232.
121. Seroussi, E., et al., *Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs*. BMC Genomics, 2010. **11**: p. 673.
122. Cicconardi, F., et al., *Massive screening of copy number population-scale variation in Bos taurus genome*. BMC Genomics, 2013. **14**(1): p. 124.
123. Hou, Y., et al., *Genomic characteristics of cattle copy number variations*. BMC Genomics, 2011. **12**: p. 127.
124. Jiang, L., et al., *Genome-wide identification of copy number variations in Chinese Holstein*. PLoS One, 2012. **7**(11): p. e48732.

125. Kadri, N.K., P.D. Koks, and T.H. Meuwissen, *Prediction of a deletion copy number variant by a dense SNP panel*. *Genet Sel Evol*, 2012. **44**: p. 7.
126. Hou, Y., et al., *Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array*. *BMC Genomics*, 2012. **13**(1): p. 376.
127. Jiang, L., et al., *Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins*. *BMC Genomics*, 2013. **14**(1): p. 131.
128. Eck, S.H., et al., *Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery*. *Genome Biol*, 2009. **10**(8): p. R82.
129. Kawahara-Miki, R., et al., *Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi*. *BMC Genomics*, 2011. **12**: p. 103.
130. Larkin, D.M., et al., *Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle*. *Proc Natl Acad Sci U S A*, 2012. **109**(20): p. 7693-8.
131. Stothard, P., et al., *Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery*. *BMC Genomics*, 2011. **12**: p. 559.
132. Zhan, B., et al., *Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping*. *BMC Genomics*, 2011. **12**: p. 557.
133. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. *Bioinformatics*, 2009. **25**(16): p. 2078-9.
134. You, N., et al., *SNP calling using genotype model selection on high-throughput sequencing data*. *Bioinformatics*, 2012. **28**(5): p. 643-50.
135. Ye, K., et al., *Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads*. *Bioinformatics*, 2009. **25**(21): p. 2865-71.

136. Liu, G.E., et al., *Assessment of genome integrity with array CGH in cattle transgenic cell lines produced by homologous recombination and somatic cell cloning*. *Genome Integr*, 2011. **2**(1): p. 6.
137. Hou, Y., et al., *Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle*. *Funct Integr Genomics*, 2012. **12**(1): p. 81-92.
138. Hester, S.D., et al., *Comparison of comparative genomic hybridization technologies across microarray platforms*. *J Biomol Tech*, 2009. **20**(2): p. 135-51.
139. Wernersson, R. and H.B. Nielsen, *OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes*. *Nucleic Acids Res*, 2005. **33**(Web Server issue): p. W611-5.
140. Wernersson, R., A.S. Juncker, and H.B. Nielsen, *Probe selection for DNA microarrays using OligoWiz*. *Nat Protoc*, 2007. **2**(11): p. 2677-91.
141. Kent, W.J., *BLAT--the BLAST-like alignment tool*. *Genome Res*, 2002. **12**(4): p. 656-64.
142. Untergasser, A., et al., *Primer3Plus, an enhanced web interface to Primer3*. *Nucleic Acids Res*, 2007. **35**(Web Server issue): p. W71-4.
143. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR*. *Nucleic Acids Res*, 2001. **29**(9): p. e45.
144. Liu, G.E., et al., *Analysis of recent segmental duplications in the bovine genome*. *BMC Genomics*, 2009. **10**: p. 571.
145. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data*. *Nucleic Acids Res*, 2010. **38**(16): p. e164.
146. Hu, Z.L., E.R. Fritz, and J.M. Reecy, *AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond*. *Nucleic Acids Res*, 2007. **35**(Database issue): p. D604-9.

147. Hu, Z.L. and J.M. Reecy, *Animal QTLdb: beyond a repository. A public platform for QTL comparisons and integration with diverse types of structural genomic information*. Mamm Genome, 2007. **18**(1): p. 1-4.
148. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nat Protoc, 2009. **4**(1): p. 44-57.
149. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*. Nucleic Acids Res, 2009. **37**(1): p. 1-13.
150. Sturn, A., J. Quackenbush, and Z. Trajanoski, *Genesis: cluster analysis of microarray data*. Bioinformatics, 2002. **18**(1): p. 207-8.
151. Gautier, M., et al., *A whole genome Bayesian scan for adaptive genetic divergence in West African cattle*. BMC Genomics, 2009. **10**: p. 550.
152. Chen, W.K., et al., *Mapping DNA structural variation in dogs*. Genome Res, 2009. **19**(3): p. 500-9.
153. Nicholas, T.J., et al., *A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog*. BMC Genomics, 2011. **12**: p. 414.
154. Nicholas, T.J., et al., *The genomic architecture of segmental duplications and associated copy number variants in dogs*. Genome Res, 2009. **19**(3): p. 491-9.
155. Doan, R., et al., *Identification of copy number variants in horses*. Genome Res, 2012. **22**(5): p. 899-907.
156. Dupuis, M.C., et al., *Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy*. Anim Genet, 2012.
157. Yalcin, B., et al., *The fine-scale architecture of structural variants in 17 mouse genomes*. Genome Biol, 2012. **13**(3): p. R18.

158. Cahan, P., et al., *The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells*. Nat Genet, 2009. **41**(4): p. 430-7.
159. Cutler, G., et al., *Significant gene content variation characterizes the genomes of inbred mouse strains*. Genome Res, 2007. **17**(12): p. 1743-54.
160. Graubert, T.A., et al., *A high-resolution map of segmental DNA copy number variation in the mouse genome*. PLoS Genet, 2007. **3**(1): p. e3.
161. Quinlan, A.R., et al., *Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome*. Genome Res, 2010. **20**(5): p. 623-35.
162. de Smith, A.J., et al., *Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases*. Hum Mol Genet, 2007. **16**(23): p. 2783-94.
163. Subramanian, H., et al., *Mas-related gene X2 (MrgX2) is a novel G protein-coupled receptor for the antimicrobial peptide LL-37 in human mast cells: resistance to receptor phosphorylation, desensitization, and internalization*. J Biol Chem, 2011. **286**(52): p. 44739-49.
164. Hanrahan, J.P., et al., *Mutations in the genes for oocyte-derived growth factors GDF9 and BMP15 are associated with both increased ovulation rate and sterility in Cambridge and Belclare sheep (Ovis aries)*. Biol Reprod, 2004. **70**(4): p. 900-9.
165. Awano, T., et al., *Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis*. Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2794-9.
166. Georges, M., et al., *Microsatellite mapping of a gene affecting horn development in Bos taurus*. Nat Genet, 1993. **4**(2): p. 206-10.
167. Schmutz, S.M., et al., *DNA marker-assisted selection of the polled condition in Charolais cattle*. Mamm Genome, 1995. **6**(10): p. 710-3.
168. Seichter, D., et al., *SNP-based association mapping of the polled gene in divergent cattle breeds*. Anim Genet, 2012. **43**(5): p. 595-8.

169. Chu, M.X., et al., *Mutations in BMPR-1B and BMP-15 genes are associated with litter size in Small Tailed Han sheep (Ovis aries)*. J Anim Sci, 2007. **85**(3): p. 598-603.
170. Luna-Nevarez, P., et al., *Single nucleotide polymorphisms in the growth hormone-insulin-like growth factor axis in straightbred and crossbred Angus, Brahman, and Romosinuano heifers: population genetic analyses and association of genotypes with reproductive phenotypes*. J Anim Sci, 2011. **89**(4): p. 926-34.
171. Rokouei, M., et al., *Monitoring inbreeding trends and inbreeding depression for economically important traits of Holstein cattle in Iran*. J Dairy Sci, 2010. **93**(7): p. 3294-302.
172. Sargolzaei, M., et al., *Extent of linkage disequilibrium in Holstein cattle in North America*. J Dairy Sci, 2008. **91**(5): p. 2106-17.
173. Stachowicz, K., et al., *Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle*. J Dairy Sci, 2011. **94**(10): p. 5160-75.
174. USDA-APHIS-VS, *The Cost of Johne's Disease to Dairy Producers*, 2007, IDEXX Laboratories.
175. Chirase, N.K. and L.W. Greene, *Dietary zinc and manganese sources administered from the fetal stage onwards affect immune response of transit stressed and virus infected offspring steer calves*. Animal Feed Science and Technology, 2001. **93**(3-4): p. 217-228.
176. Womack, J.E., *Bovine genomics*, 2012, Wiley-Blackwell, : Ames, Iowa. p. 1 online resource.
177. Barris, W., et al., *Next generation sequencing of African and Indicine cattle to identify single nucleotide polymorphisms*. Animal Production Science, 2012. **52**(2-3): p. 133-142.
178. Canavez, F.C., et al., *Genome sequence and assembly of Bos indicus*. Journal of Heredity, 2012. **103**(3): p. 342-348.

179. Alkan, C., B.P. Coe, and E.E. Eichler, *APPLICATIONS OF NEXT-GENERATION SEQUENCING Genome structural variation discovery and genotyping*. Nature Reviews Genetics, 2011. **12**(5): p. 363-375.
180. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences*. Nucleic Acids Res, 1999. **27**(2): p. 573-80.
181. Goujon, M., et al., *A new bioinformatics analysis tools framework at EMBL-EBI*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W695-9.
182. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0*. Bioinformatics, 2007. **23**(21): p. 2947-8.
183. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*. Genome Biol, 2010. **11**(8): p. R86.
184. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists*. Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.
185. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
186. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood*. Mol Biol Evol, 2007. **24**(8): p. 1586-91.
187. Fontanesi, L., E. Scotti, and V. Russo, *Haplotype variability in the bovine MITF gene and association with piebaldism in Holstein and Simmental cattle breeds*. Anim Genet, 2012. **43**(3): p. 250-6.
188. Grisart, B., et al., *Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition*. Genome Res, 2002. **12**(2): p. 222-31.
189. Doan, R., et al., *Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare*. BMC Genomics, 2012. **13**: p. 78.

190. Zhang, J., *Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes*. J Mol Evol, 2000. **50**(1): p. 56-68.
191. Fujimoto, A., et al., *Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing*. Nat Genet, 2010. **42**(11): p. 931-6.
192. Gautier, M., et al., *Characterization of the DGAT1 K232A and variable number of tandem repeat polymorphisms in French dairy cattle*. J Dairy Sci, 2007. **90**(6): p. 2980-8.
193. Grisart, B., et al., *Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition*. Proc Natl Acad Sci U S A, 2004. **101**(8): p. 2398-403.
194. Lacorte, G.A., et al., *DGAT1 K232A polymorphism in Brazilian cattle breeds*. Genet Mol Res, 2006. **5**(3): p. 475-82.
195. Kaupe, B., et al., *DGAT1 polymorphism in Bos indicus and Bos taurus cattle breeds*. J Dairy Res, 2004. **71**(2): p. 182-7.
196. Zheng, W., et al., *Regulatory variation within and between species*. Annu Rev Genomics Hum Genet, 2011. **12**: p. 327-46.
197. Kasowski, M., et al., *Variation in transcription factor binding among humans*. Science, 2010. **328**(5975): p. 232-5.
198. Schmidt, D., et al., *Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding*. Science, 2010. **328**(5981): p. 1036-40.
199. *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution*. Nature, 2004. **432**(7018): p. 695-716.
200. Wagschal, A., et al., *Chromatin Immunoprecipitation (ChIP) on Unfixed Chromatin from Cells and Tissues to Analyze Histone Modifications*. CSH Protoc, 2007. **2007**: p. pdb prot4767.

201. Pekowska, A., et al., *H3K4 tri-methylation provides an epigenetic signature of active enhancers*. EMBO J, 2011. **30**(20): p. 4198-210.
202. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences*. Nature, 2009. **462**(7271): p. 315-22.
203. Chodavarapu, R.K., et al., *Relationship between nucleosome positioning and DNA methylation*. Nature, 2010. **466**(7304): p. 388-92.
204. VanRaden, P.M., et al., *Genomic imputation and evaluation using high-density Holstein genotypes*. J Dairy Sci, 2013. **96**(1): p. 668-78.
205. Hou, Y., et al., *Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake*. Funct Integr Genomics, 2012. **12**(4): p. 717-23.
206. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet, 2007. **81**(3): p. 559-75.
207. Cantsilieris, S. and S.J. White, *Correlating multiallelic copy number polymorphisms with disease susceptibility*. Hum Mutat, 2013. **34**(1): p. 1-13.
208. Bentley, R.W., et al., *Association of higher DEFB4 genomic copy number with Crohn's disease*. Am J Gastroenterol, 2010. **105**(2): p. 354-9.
209. Hollox, E.J., et al., *Psoriasis is associated with increased beta-defensin genomic copy number*. Nat Genet, 2008. **40**(1): p. 23-5.
210. Fellermann, K., et al., *A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon*. Am J Hum Genet, 2006. **79**(3): p. 439-48.
211. Traherne, J.A., *Human MHC architecture and evolution: implications for disease association studies*. Int J Immunogenet, 2008. **35**(3): p. 179-92.
212. Olsson, L.M. and R. Holmdahl, *Copy number variation in autoimmunity--importance hidden in complexity?* Eur J Immunol, 2012. **42**(8): p. 1969-76.

213. Chang, F.M., et al., *The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus*. Hum Genet, 1996. **98**(1): p. 91-101.
214. Durr, U.H., U.S. Sudheendra, and A. Ramamoorthy, *LL-37, the only human member of the cathelicidin family of antimicrobial peptides*. Biochim Biophys Acta, 2006. **1758**(9): p. 1408-25.
215. Braff, M.H., et al., *Keratinocyte production of cathelicidin provides direct activity against bacterial skin pathogens*. Infect Immun, 2005. **73**(10): p. 6771-81.
216. Bals, R. and P.S. Hiemstra, *Innate immunity in the lung: how epithelial cells fight against respiratory pathogens*. Eur Respir J, 2004. **23**(2): p. 327-33.
217. Morizane, S., et al., *Kallikrein expression and cathelicidin processing are independently controlled in keratinocytes by calcium, vitamin D(3), and retinoic acid*. J Invest Dermatol, 2010. **130**(5): p. 1297-306.
218. Gombart, A.F., T. Saito, and H.P. Koefler, *Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates*. BMC Genomics, 2009. **10**: p. 321.
219. Peric, M., et al., *VDR and MEK-ERK dependent induction of the antimicrobial peptide cathelicidin in keratinocytes by lithocholic acid*. Mol Immunol, 2009. **46**(16): p. 3183-7.
220. D'Aldebert, E., et al., *Bile salts control the antimicrobial peptide cathelicidin through nuclear receptors in the human biliary epithelium*. Gastroenterology, 2009. **136**(4): p. 1435-43.
221. Hancock, R.E. and G. Diamond, *The role of cationic antimicrobial peptides in innate host defences*. Trends Microbiol, 2000. **8**(9): p. 402-10.
222. Park, K., et al., *Regulation of cathelicidin antimicrobial peptide expression by an endoplasmic reticulum (ER) stress signaling, vitamin D receptor-independent pathway*. J Biol Chem, 2011. **286**(39): p. 34121-30.
223. Harder, J., et al., *Differential gene induction of human beta-defensins (hBD-1, -2, -3, and -4) in keratinocytes is inhibited by retinoic acid*. J Invest Dermatol, 2004. **123**(3): p. 522-9.

224. Nelson, C.D., et al., *Vitamin D signaling in the bovine immune system: a model for understanding human vitamin D requirements*. *Nutrients*, 2012. **4**(3): p. 181-96.
225. Nelson, C.D., et al., *Modulation of the bovine innate immune response by production of 1alpha,25-dihydroxyvitamin D(3) in bovine monocytes*. *J Dairy Sci*, 2010. **93**(3): p. 1041-9.
226. Mossallam, A.A.A., *Expression of different isoforms of Cathelicidin-4 transcripts in river Buffalo mRNA*. *Journal of Applied Biosciences*, 2008. **6**: p. 150-157.
227. Mossallam, A.A.A., et al., *Retention of Intron-1 in Cathelicidin-4 mRNA of Egyptian Native and Frisian Crossbred Cattle*. *Journal of Applied Sciences Research*, 2007. **3**(11): p. 1400-1406.
228. Harford, J.B., *Preparation and isolation of cells*, in *Current Protocols in Cell Biology* 2001, John Wiley & Sons, Inc.

APPENDIX 2.1

DNA ISOLATION PROTOCOL

WBC Isolation from Blood:

- Remove serum from blood (clear layer)
- Place ~500 µl of remaining blood into 2 mL tubes
- Add 1 mL red blood cell lysis (RBC) solution to each tube
- Rotate at room temperature (RT) for 3 minutes
- Centrifuge at RT for 1 minute at 17,000xg
- Pour off liquid (ensuring pellet remains in tube)
- Add 1mL red blood cell lysis (RBC) solution to each tube
- Rotate at RT for 3 minutes
- Centrifuge at RT for 1 minute at 17,000xg
- Pour off liquid (ensuring pellet remains in tube)
- Freeze at -80°C or proceed to DNA isolation

DNA Isolation:

- Add 500 µl of NTES and 1 µl of Proteinase K to isolated WBC (~400 µl) or small amount of tissue in 1.5 ml tube
- Wrap top of tubes with parafilm
- Rock overnight at 55°C
- Add 500 µl PCI (Phenyl-Chloroform-Isoamyl) to the tube containing digested DNA.
- Rock at room temp for 10 minutes
- Centrifuge samples at 12000xg for 10 minutes
- Prepare Phase Lock tubes by centrifugation at 12000xg for 1 minute
- Carefully remove top layer without disturbing the bottom layer.
- Place top layer into phase lock tubes
- Add 500 µl PCI
- Rock 10 min at room temp.
- Spin at 12000xg for 5 minutes

- (If phase tubes are full, remove top layer of liquid and place into new phase tube before continuing)
- Add 500 μ l Chloroform
- Rock 10 minutes at room temp.
- Spin at 12000xg for 5 minutes
- Remove top aqueous layer and place into new 1.5 ml or 2 ml (depending on volume) tube
- Add 850 μ l 100% Isopropanol
- Mix by rocking in hand until DNA is visible and forms a loose ball (may have to vortex gently or add more isopropanol if using larger volumes)
- Spin at 18000xg for 3 minutes
- Remove all liquid by pouring or pipetting, while ensuring that DNA pellet remains in tube
- Add 1 ml of 70% ethanol
- Rock at least 30 minutes, if DNA appears any color other than white, rock longer possibly overnight.
- Spin at 18000xg for 3 min
- Remove excess ethanol, while ensuring pellet remains in tube
- Place tube with caps open on heat block set at 37°C (up to 55°C) until all ethanol is evaporated.
- Re-suspend in Elution Buffer (EB) (any manufacturer) (usually 200 μ l work but this will be adjusted based on the size of pellet. If unsure start small (100 μ l) and add more EB later, if the DNA is too thick.
- Place on heat block at 37°C for several hours to ensure that the DNA is in solution (may vortex gently throughout process)
- DNA should now be pure and of concentrations at least 500 ng/ μ l (depending on starting volumes)

APPENDIX 2.2

ARRAY DESIGN USING OLIGOWIZ2.0

Selection of sequence for oligonucleotides by two methods:

1. UCSC Genome Browser:

- Determine type of region to include on array (gene, exons, intergenic, etc.)
 - If exonic: (typically ensembl) – go to UCSC Genome browser:
<http://genome.ucsc.edu/>
 - Select **Tables** from menu button along top of screen
 - A new screen will appear. (See image below)
 - Within screen select:
 - Genome = Cow (change if wanting a different species)
 - Assembly = Oct. 2007 Baylor 4.0 (change for different assembly)
 - Group = Genes and Gene Prediction Tracks
 - Track = Ensembl Genes (change for different annotation)
 - Region = position = chrXX:start-end to download all genes on chr
 - Output format = sequence
 - Output file: Type in a name for your fasta formatted text file
 - Click: “get output” button at bottom of page

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the [OpenHelix Table Browser tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: **genome:** **assembly:**

group: **track:**

table:

region: genome position

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: Send output to [Galaxy](#) [GREAT](#)

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

- A new screen will appear. (See image below)

Select sequence type for Ensembl Genes

genomic
 protein

- Select genomic
- Click 'submit'
- A new window will appear (See image below)
 - Check the boxes for genic elements to be covered by array
 - For exome array: select:
 - 5' UTR Exons
 - 3' UTR Exons
 - CDS Exons
 - One fasta record per region
 - Add 40 bp upstream and downstream
 - All Upper case
 - Click 'get sequence' to begin file download

Ensembl Genes Genomic Sequence

Sequence Retrieval Region Options:

Promoter/Upstream by bases

5' UTR Exons

CDS Exons

3' UTR Exons

Introns

Downstream by bases

One FASTA record per gene.

One FASTA record per region (exon, intron, etc.) with extra bases upstream (5') and extra downstream (3')

Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

Exons in upper case, everything else in lower case.

CDS in upper case, UTR in lower case.

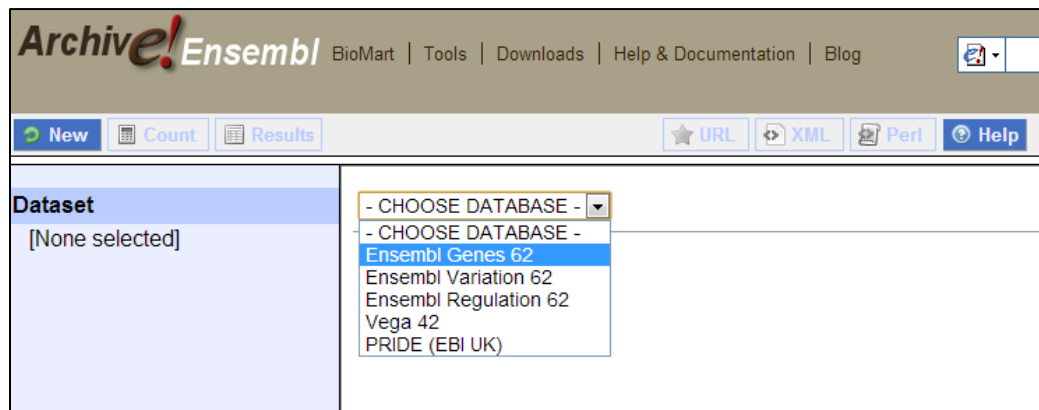
All upper case.

All lower case.

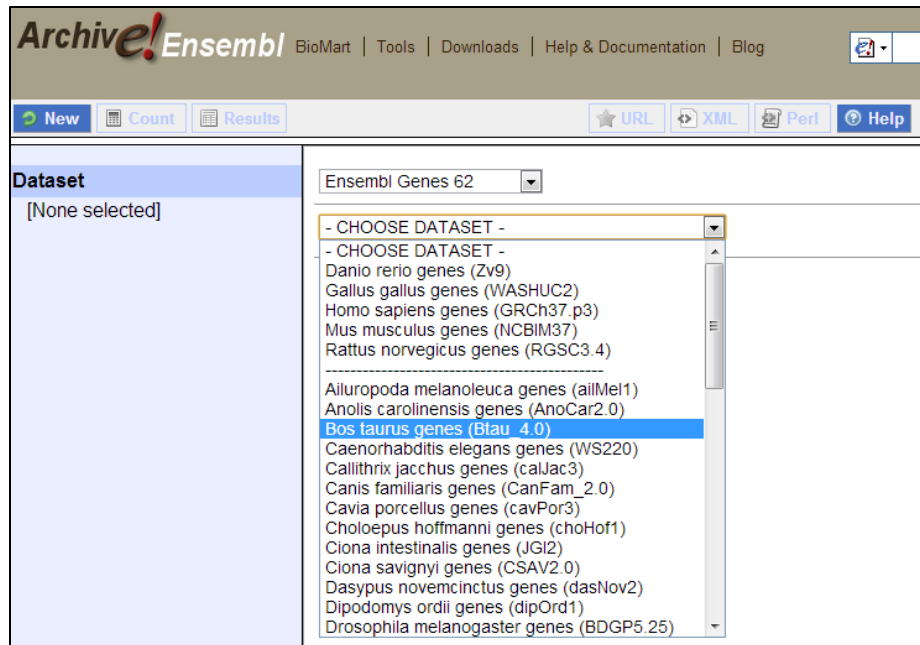
Mask repeats: to lower case to N

2. Ensembl BioMart:

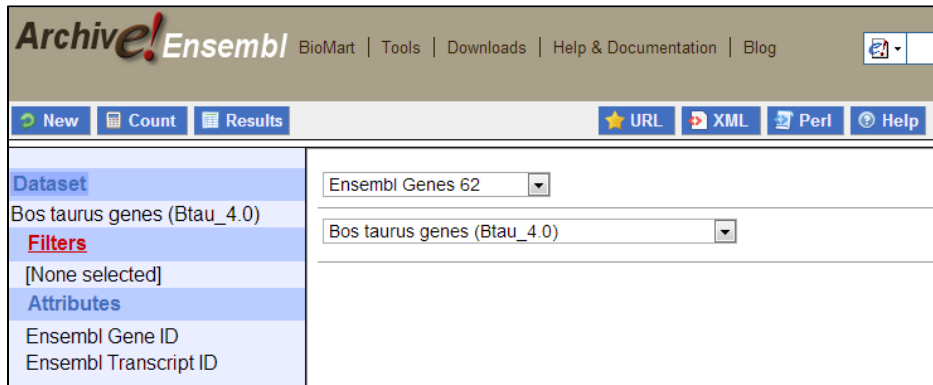
- Determine type of region to include on array (gene, exons, intergenic, etc.)
 - If exonic: (typically ensembl) – go to Ensembl BioMart:
 - <http://uswest.ensembl.org/biomart/martview> (for older annotations, including bosTau4.0, use archive site:
 - <http://apr2011.archive.ensembl.org/biomart/martview/>
 - Select Ensembl Genes 62 (or newest version available if not using archive)



- Select Species: Bos taurus genes (Btau 4.0)



- Select Filters from Options along left side (See image below)



- Select GENE (See image below)
 - Under Gene Type, select types of genes (usually exclude pseudo and retro genes)

Please restrict your query using criteria below

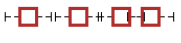
Dataset Bos taurus genes (Btau_4.0) Filters [None selected] Attributes Ensembl Gene ID Ensembl Transcript ID <hr/> Dataset [None Selected]	<input type="checkbox"/> REGION: <input type="checkbox"/> GENE: <input type="checkbox"/> Limit to genes ... with WikiGene ID(s) <input checked="" type="radio"/> Only <input type="radio"/> Excluded <input type="checkbox"/> ID list limit Ensembl Gene ID(s) [e.g. ENSG00000139618] <input type="text"/> <input type="button" value="Choose File"/> No file chosen <input type="checkbox"/> Transcript count >= <input type="text"/> <input type="checkbox"/> Gene type miRNA misc_RNA Mt_rRNA Mt_tRNA protein_coding
--	--

- Select REGION (See image below)
 - Under Region, select chromosome and start-end of chromosome

Please restrict your query using criteria below

<input type="checkbox"/> REGION:	
<input type="checkbox"/> Chromosome	1 <input type="button" value="v"/>
<input type="checkbox"/> Base pair	
Gene Start (bp)	1 <input type="text"/>
Gene End (bp)	10000000 <input type="text"/>
<input type="checkbox"/> Marker	
Marker Start	<input type="text"/>
Marker End	<input type="text"/>
<input checked="" type="checkbox"/> Multiple Chromosomal Regions (Chr:Start:End:Strand)	
Chromosome Regions	<input type="text"/>
<input type="button" value="Choose File"/> No file chosen	

- Select Attributes from options along left side (See image below)
 - Select Sequences
 - Select Exon sequences (for exome array)
 - Add 40bp of flanking sequence upstream AND downstream

Dataset Bos taurus genes (Btau_4.0)	<input type="radio"/> Features <input type="radio"/> Variation <input type="radio"/> Structures <input checked="" type="radio"/> Sequences <input type="radio"/> Homologs
Filters [None selected]	SEQUENCES: Sequences (max 1)
Attributes Ensembl Gene ID Ensembl Transcript ID Exon sequences Upstream flank [40] Downstream flank [40]	
Dataset [None Selected]	<input type="radio"/> Unspliced (Transcript) <input type="radio"/> 5' UTR <input type="radio"/> Unspliced (Gene) <input type="radio"/> 3' UTR <input type="radio"/> Flank (Transcript) <input checked="" type="radio"/> Exon sequences <input type="radio"/> Flank (Gene) <input type="radio"/> cDNA sequences <input type="radio"/> Flank-coding region (Transcript) <input type="radio"/> Coding sequence <input type="radio"/> Flank-coding region (Gene) <input type="radio"/> Peptide
	Upstream flank <input checked="" type="checkbox"/> Upstream flank [40]
	Downstream flank <input checked="" type="checkbox"/> Downstream flank [40]

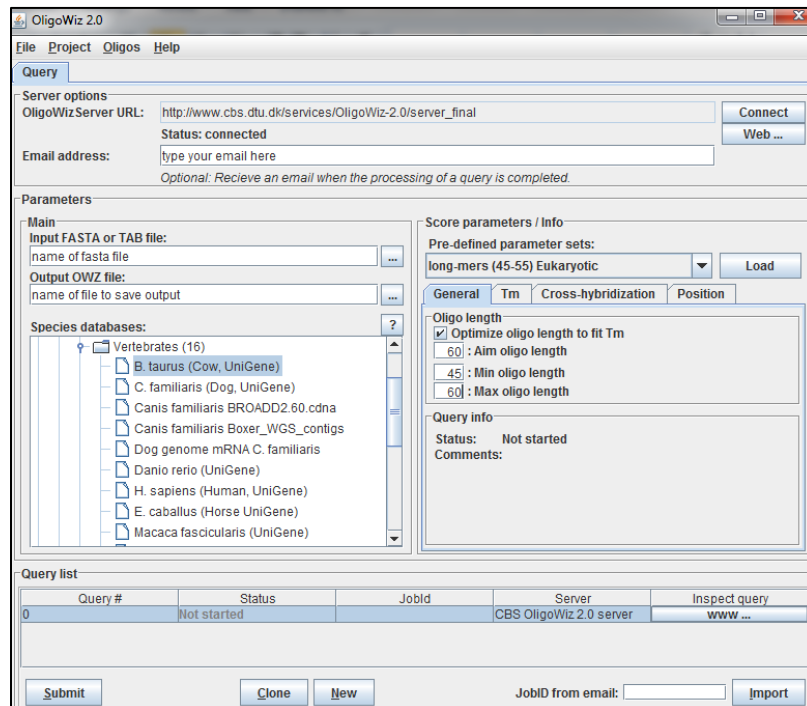
- Scroll down and select Header information (See image below)
 - Select the following boxes:
 - Ensembl Gene ID
 - Ensembl Exon ID
 - Chromosome Name
 - Gene Start
 - Gene End
 - Ensembl Exon Start
 - Ensembl Exon End
 - Select Results and save fasta file

Dataset Bos taurus genes (Btau_4.0)	Header Information
Filters [None selected]	Gene Information <input checked="" type="checkbox"/> Ensembl Gene ID <input checked="" type="checkbox"/> Chromosome Name <input type="checkbox"/> Description <input type="checkbox"/> Gene Start (bp) <input type="checkbox"/> Associated Gene Name <input type="checkbox"/> Gene End (bp) <input type="checkbox"/> Associated Gene DB <input type="checkbox"/> Ensembl Protein Family ID(s)
Attributes Ensembl Gene ID Exon sequences Upstream flank [40] Downstream flank [40] Chromosome Name Ensembl Exon ID Exon Chr Start (bp) Exon Chr End (bp)	Transcript Information <input type="checkbox"/> CDS start (within cDNA) <input type="checkbox"/> Ensembl Transcript ID <input type="checkbox"/> CDS end (within cDNA) <input type="checkbox"/> Ensembl Protein ID <input type="checkbox"/> 5' UTR Start <input type="checkbox"/> Strand <input type="checkbox"/> 5' UTR End <input type="checkbox"/> Transcript Start (bp) <input type="checkbox"/> 3' UTR Start <input type="checkbox"/> Transcript End (bp) <input type="checkbox"/> 3' UTR End
Dataset [None Selected]	Exon Information <input type="checkbox"/> CDS Length <input checked="" type="checkbox"/> Exon Chr End (bp) <input type="checkbox"/> CDS Start <input type="checkbox"/> Strand <input type="checkbox"/> CDS End <input type="checkbox"/> Exon Rank in Transcript <input checked="" type="checkbox"/> Ensembl Exon ID <input type="checkbox"/> Constitutive Exon <input checked="" type="checkbox"/> Exon Chr Start (bp)

OligoWiz Probe Selection:

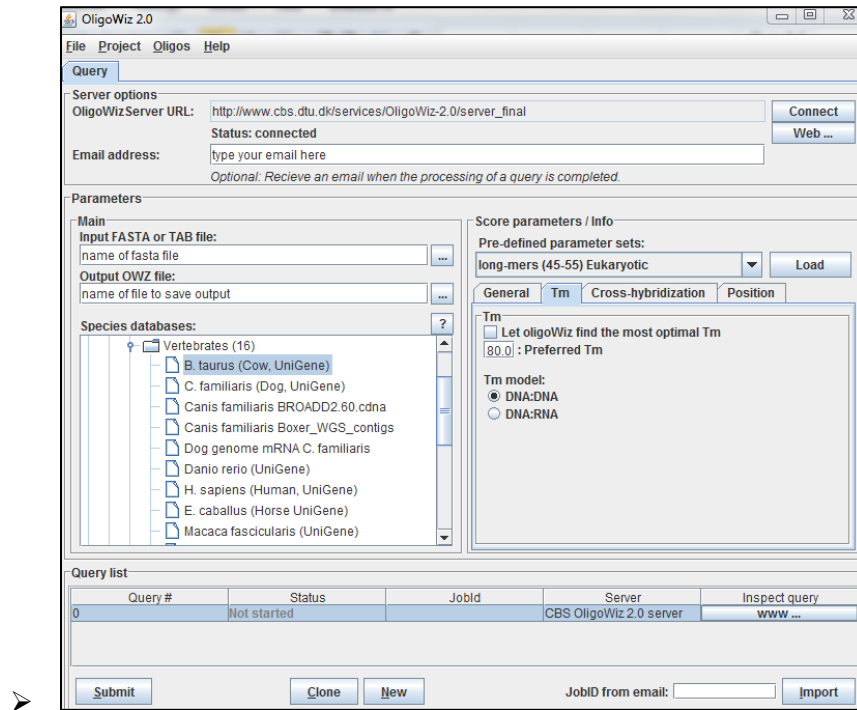
- Download current version of OligoWiz (protocol based on version 2.1.3)

- <http://www.cbs.dtu.dk/services/OligoWiz/>
- Will need to have newest version of java installed:
 - <http://www.java.com/en/>
- Open OligoWiz program (no installation) (See image below)
- Enter your email address, you will receive email with link to download your data if the program is closed
- Select you input fasta file
- Select location and name of the output file to be created by OligoWiz
- Select reference from drop-down list – B. Taurus Unigene or Whole genome
- Select: long-mers (45-55) Eukaryotic
- General settings:
 - Check: Optimize oligo length to fit Tm
 - Aim Oligo Length = 60bp
 - Min Oligo Length = 45bp
 - Max Oligo Length = 60bp

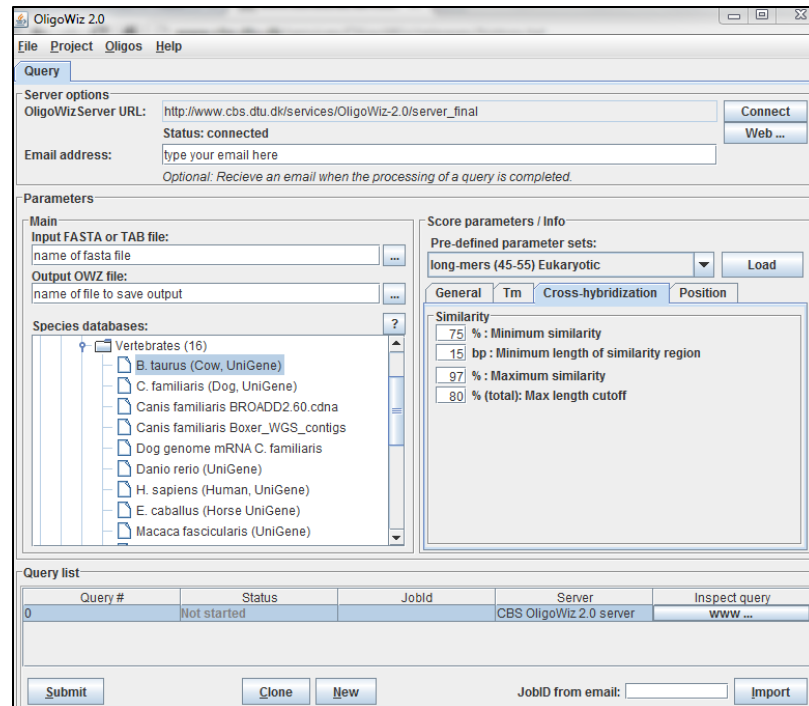


- Select Tm Tab (See image below)
 - Uncheck: Let OligoWiz find the most optimal Tm
 - Preferred Tm = 80.0

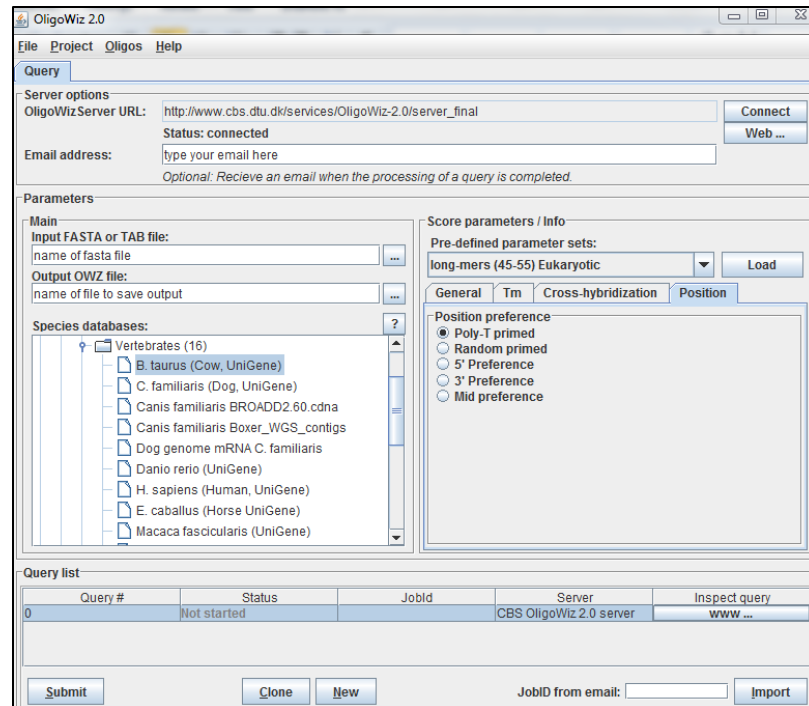
- Tm model = DNA:DNA



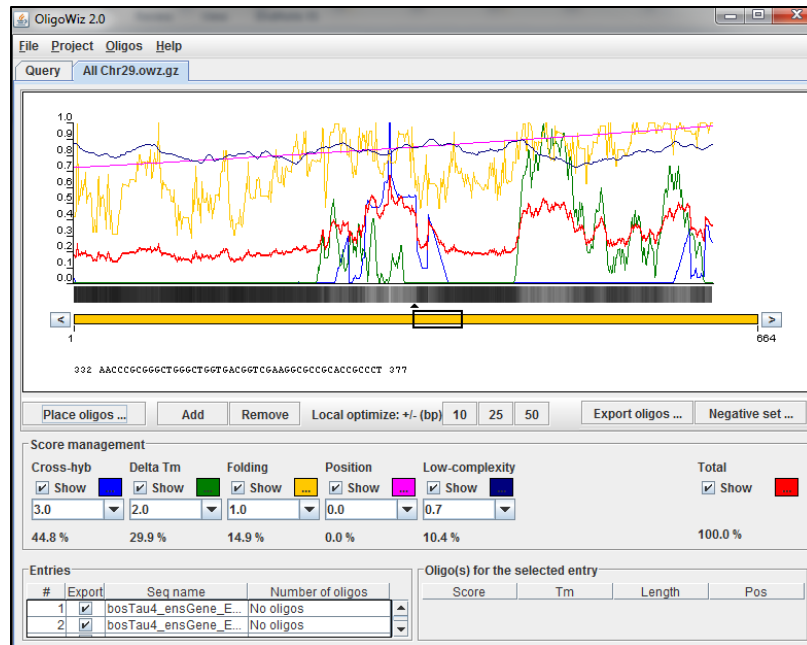
- Select: Cross Hybridization Tab (See image below)
 - Maximum similarity = 75%
 - Minimum length of similarity region = 15bp
 - Maximum similarity = 97%
 - % total Max Length Cutoff = 80%



- Select Position tab (See image below)
 - Position preference = Poly-T primed
 - Click Submit

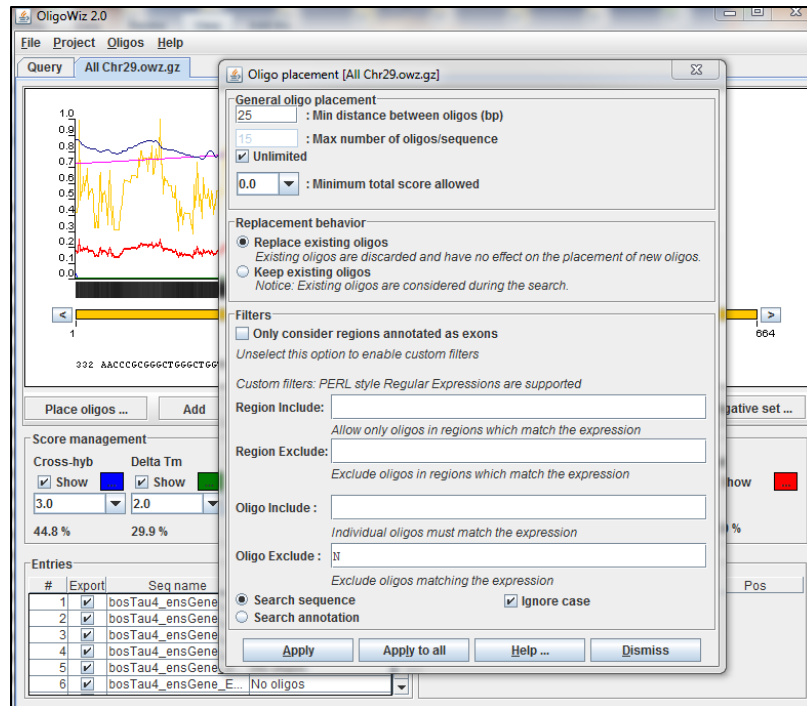


- Once the job has been completed, Status will be Complete
 - Select completed query from Query List to open results (See image below)
 - Alternatively, an email will provide a link to download the results
 - Download and open oligowiz
 - File – Open OWZ data file – select file and open (See image below)
 - Select Score management settings:
 - Cross-Hyb = 3.0
 - Delta Tm = 2.0
 - Folding = 1.0
 - Position = 0.0
 - Low-complexity = 0.7

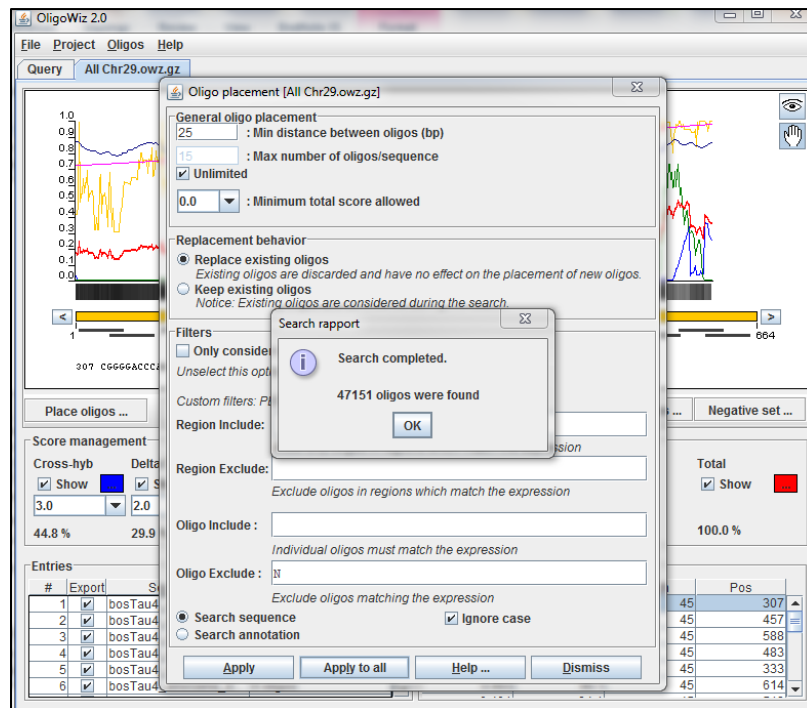


➤ Select Place Oligos (See image below)

- Min distance between oligos = 25 bp (allows ~half of a probe to overlap another probe (do not go below this for 45-60bp probes))
- Check Unlimited
- Minimum total score = 0.0 (allows for selection of all probes)
- Uncheck: Only consider regions annotated as exons
- Oligo Exclude = N
- Select: Search Sequences
- Check: Ignore case
- Click: Apply to All to predict probes for all fasta sequences

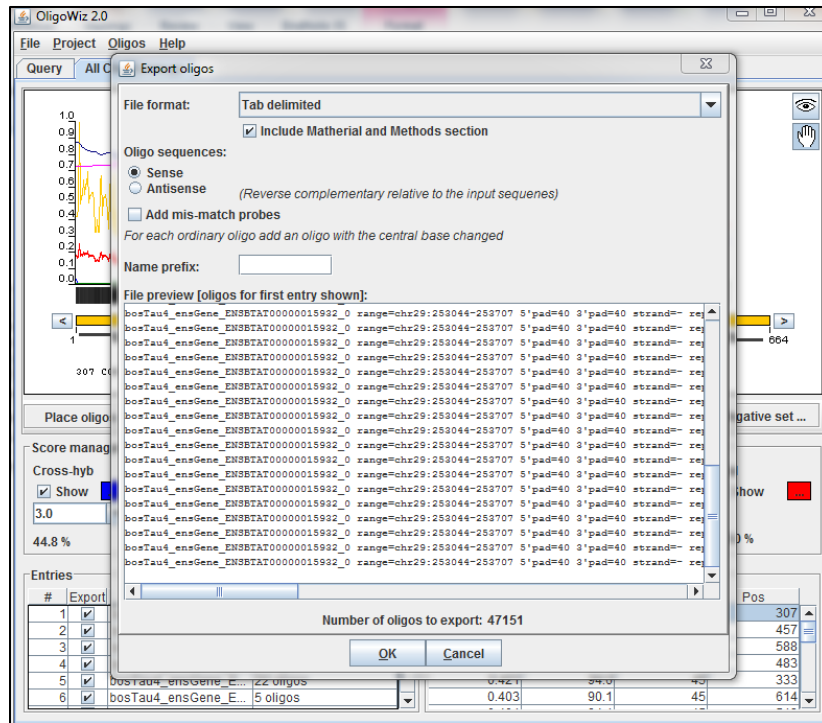


- Once search is completed a pop-up window will state “Search Completed”
 - Click OK



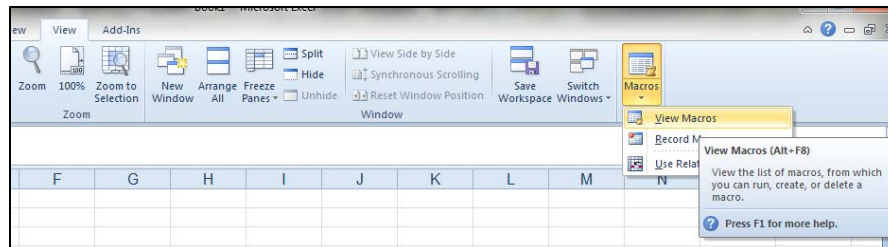
- Select Export Oligos (See image below)

- Select
 - Tab delimited
 - Sense
 - Include material and methods section
 - Click OK to save probe file

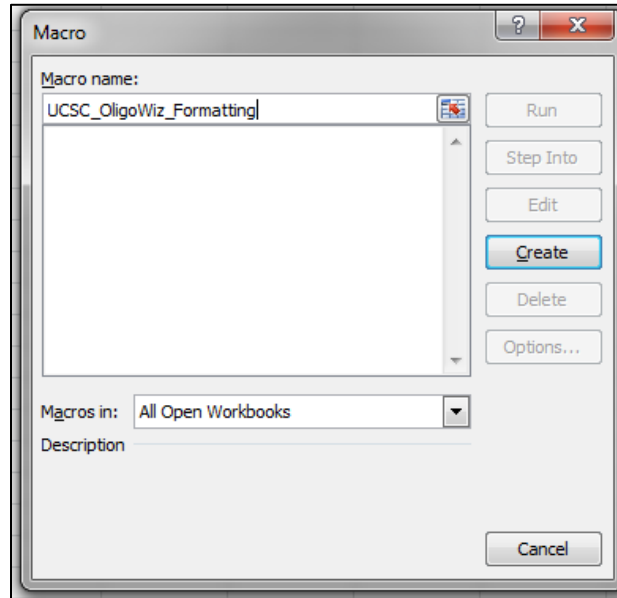


- Exported file will be a tab delimited file
 - Open file within Microsoft excel for custom sorting and filtering
- Custom Probe Sorting and Selection:**
- Open oligo file in excel (only use tab delimiters when opening)

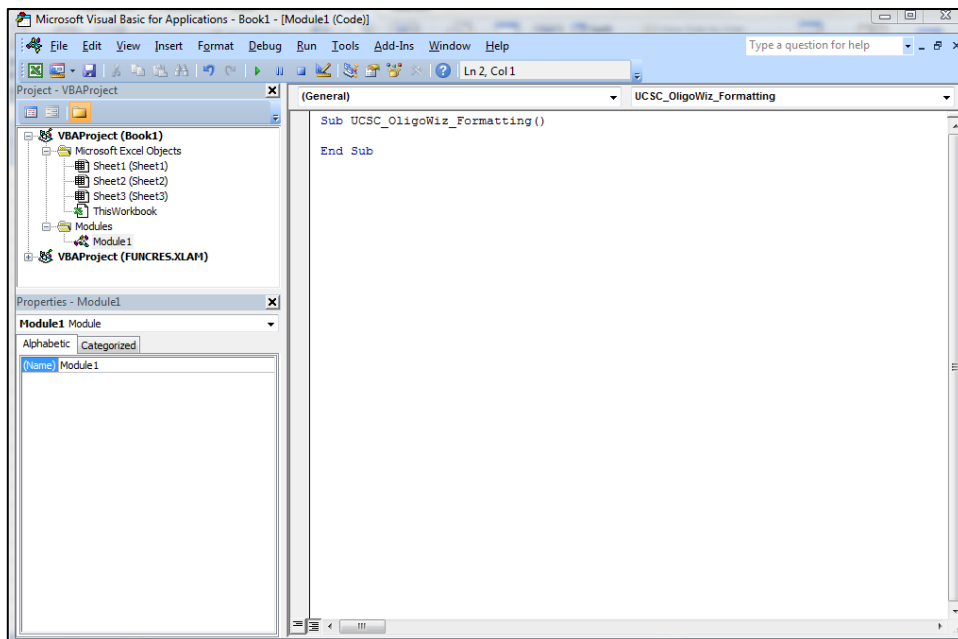
- Select the View tab (See image below)
 - Select Macros
 - Click View Macros



- A new window will open (See image below)
 - Type the name of the new macro
 - Click on Create

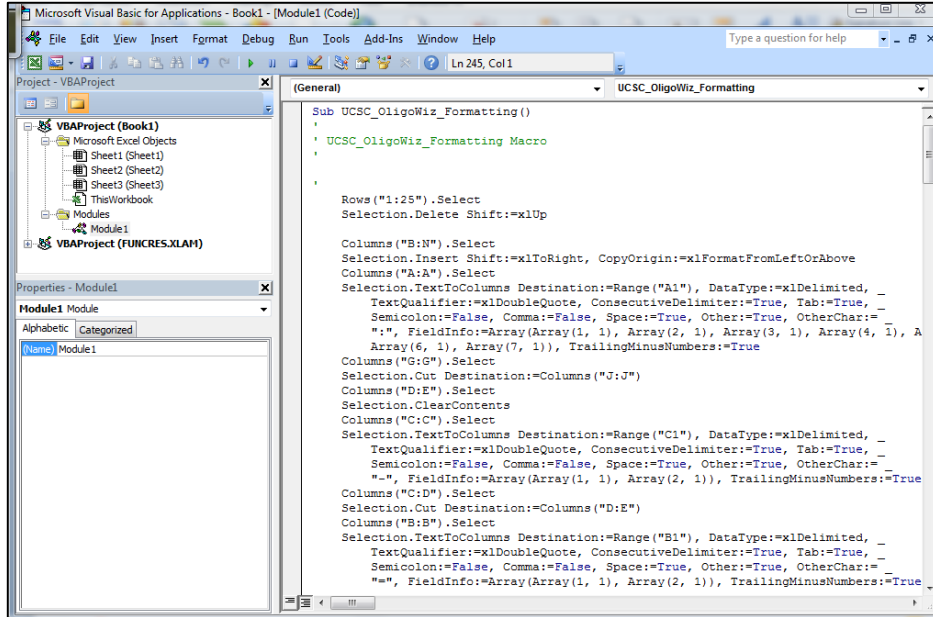


- A new window will open (See image below)
 - Copy and Paste the entire macro into window (see macro below)
 - Be sure to include everything from Sub to End Sub
 - Replace existing empty Macro text (Sub to End Sub)



- Once the macro is pasted you should see a complete macro (See image below)

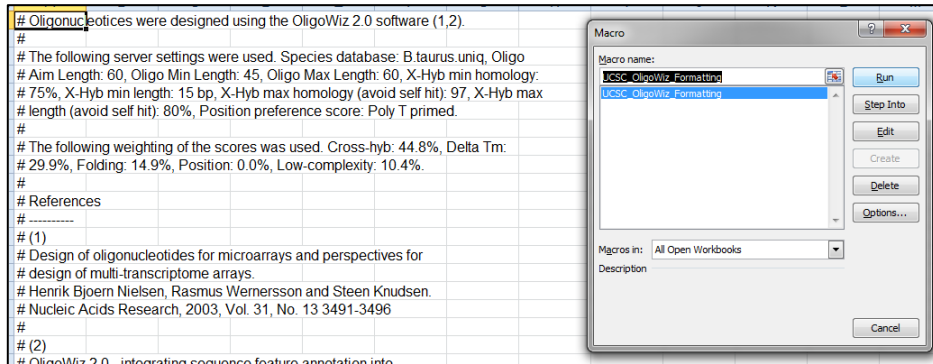
- You can close the entire visual basic screen (no saving required)



-

- To run the new macro
 - Go back to the macro option (where you typed the macro name and clicked create) (See image below)

- Select the macro name and click Run
 - The screen will rapidly change as it processes the file



-

Custom Macro (Copy everything below through End Sub)

```
Sub UCSC_OligoWiz_Formatting()
'
' UCSC_OligoWiz_Formatting Macro
'
'
    Rows("1:25").Select
    Selection.Delete Shift:=xlUp

    Columns("B:N").Select
    Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
    Columns("A:A").Select
    Selection.TextToColumns Destination:=Range("A1"), DataType:=xlDelimited, _
        TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
        Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
        ":", FieldInfo:=Array(Array(1, 1), Array(2, 1), Array(3, 1), Array(4, 1), Array(5, 1), _
        Array(6, 1), Array(7, 1)), TrailingMinusNumbers:=True
    Columns("G:G").Select
    Selection.Cut Destination:=Columns("J:J")
    Columns("D:E").Select
    Selection.ClearContents
    Columns("C:C").Select
    Selection.TextToColumns Destination:=Range("C1"), DataType:=xlDelimited, _
        TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
        Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
        "-", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
    Columns("C:D").Select
    Selection.Cut Destination:=Columns("D:E")
    Columns("B:B").Select
    Selection.TextToColumns Destination:=Range("B1"), DataType:=xlDelimited, _
        TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
        Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
        "=", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
    Columns("F:F").Select
    Selection.TextToColumns Destination:=Range("F1"), DataType:=xlDelimited, _
        TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
        Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
        "=", FieldInfo:=Array(1, 1), TrailingMinusNumbers:=True
    Columns("F:F").Select
    Selection.Delete Shift:=xlToLeft
    Columns("C:F").Select
    Selection.Cut Destination:=Columns("E:H")
    Columns("I:I").Select
    Selection.TextToColumns Destination:=Range("I1"), DataType:=xlDelimited, _
        TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
        Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
        "=", FieldInfo:=Array(1, 1), TrailingMinusNumbers:=True
    Columns("I:I").Select
    Selection.Delete Shift:=xlToLeft
    Selection.TextToColumns Destination:=Range("I1"), DataType:=xlDelimited, _
        TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
        Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
        "-", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
    Columns("J:J").Select
    Selection.Cut Destination:=Columns("K:K")
    Columns("I:I").Select
    Selection.TextToColumns Destination:=Range("I1"), DataType:=xlDelimited, _
```

```

TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
"_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
Selection.Delete Shift:=xlToLeft
Columns("B:B").Select
Selection.Delete Shift:=xlToLeft
Columns("B:C").Select
Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
Columns("A:A").Select
Selection.TextToColumns Destination:=Range("A1"), DataType:=xlDelimited, _
TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
Semicolon:=False, Comma:=False, Space:=True, Other:=True, OtherChar:= _
"_", FieldInfo:=Array(Array(1, 1), Array(2, 1), Array(3, 1), Array(4, 1)), _
TrailingMinusNumbers:=True
Columns("A:B").Select
Selection.Delete Shift:=xlToLeft
Range("A3").Select
Columns("A:A").EntireColumn.AutoFit
Range("C1").Select
Selection.End(xlDown).Select
Selection.End(xlUp).Select
Range("B1").Select
Selection.End(xlDown).Select
Range("B376").Select
Selection.End(xlUp).Select
Selection.End(xlToLeft).Select
Selection.End(xlUp).Select
Selection.End(xlUp).Select
Selection.End(xlUp).Select
Columns("e:f").Select
Selection.Cut
Columns("c:c").Select
Selection.Insert Shift:=xlToRight
Columns("h:i").Select
Selection.Cut
Columns("e:e").Select
Selection.Insert Shift:=xlToRight
Columns("i:i").Select
Selection.Cut
Columns("g:g").Select
Selection.Insert Shift:=xlToRight
Columns("i:i").Select
Selection.Cut
Columns("h:h").Select
Selection.Insert Shift:=xlToRight
Range("i1").Select
ActiveWindow.ScrollColumn = 2
ActiveWindow.ScrollColumn = 3
ActiveWindow.ScrollColumn = 4
ActiveWindow.ScrollColumn = 5
ActiveWindow.ScrollColumn = 6
Range("L:L,M:M,N:N,O:O,P:P,Q:Q,R:R,S:S").Select
Range("S1").Activate
Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
Columns("v:v").Select
Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
Columns("m:m").Select
Selection.TextToColumns Destination:=Range("m1"), DataType:=xlDelimited, _
TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _

```

Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Columns("n:n").Select
 Selection.TextToColumns Destination:=Range("n1"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Columns("o:o").Select
 Selection.TextToColumns Destination:=Range("o1"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Columns("p:p").Select
 Selection.TextToColumns Destination:=Range("p1"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Columns("q:q").Select
 Selection.TextToColumns Destination:=Range("q1"), DataType:=xlFixedWidth, _
 OtherChar:=" ", FieldInfo:=Array(Array(0, 1), Array(5, 1), Array(9, 1)), _
 TrailingMinusNumbers:=True
 Columns("q:r").Select
 Range("r1").Activate
 Selection.Delete Shift:=xlToLeft
 Columns("r:r").Select
 Selection.TextToColumns Destination:=Range("r1"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Columns("s:s").Select
 Selection.TextToColumns Destination:=Range("s1"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Columns("t:t").Select
 Selection.TextToColumns Destination:=Range("t1"), DataType:=xlDelimited, _
 TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
 Semicolon:=False, Comma:=False, Space:=True, Other:=False, OtherChar _
 := "_", FieldInfo:=Array(Array(1, 1), Array(2, 1)), TrailingMinusNumbers:=True
 Selection.Delete Shift:=xlToLeft
 Selection.End(xlToLeft).Select
 Columns("l:l").Select
 Selection.Delete Shift:=xlToLeft
 Range("j1").Select
 Selection.End(xlToLeft).Select
 Selection.End(xlToLeft).Select
 Selection.End(xlToLeft).Select
 Selection.End(xlToLeft).Select
 Columns("j:j").Select
 Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
 Range("i1").Select
 ActiveCell.FormulaR1C1 = "=IF(RC[-2]=""+""",RC[-6]+RC[-4]-1,RC[-5]-RC[-3]+1)"
 Range("j1").Select

```

ActiveCell.FormulaR1C1 = "=RC[-1]+RC[4]-1"
Range("k1").Select
ActiveCell.FormulaR1C1 = "=RC[-3]&""&RC[-2]&""&RC[-1]"
Range("i1:k1").Select
Range("k1").Activate
Selection.Copy
Range("h1").Select
Selection.End(xlDown).Select
Selection.End(xlUp).Select
Range("h1:k1").Select
Application.CutCopyMode = False
Selection.Copy
Range("h1").Select
Selection.End(xlDown).Select
Range(Selection, Selection.End(xlUp)).Select
ActiveSheet.Paste
Selection.End(xlUp).Select
Selection.End(xlUp).Select
Selection.End(xlUp).Select
Selection.End(xlToLeft).Select

Rows("1:1").Select
Selection.Insert Shift:=xlDown, CopyOrigin:=xlFormatFromLeftOrAbove
Range("A1").Select
ActiveCell.FormulaR1C1 = "Transcript ID"
Range("B1").Select
ActiveCell.FormulaR1C1 = "Exon #"
Range("c1").Select
ActiveCell.FormulaR1C1 = "Range Start"
Range("d1").Select
ActiveCell.FormulaR1C1 = "Range End"
Range("e1").Select
ActiveCell.FormulaR1C1 = "S"
Range("f1").Select
ActiveCell.FormulaR1C1 = "E"
Range("g1").Select
ActiveCell.FormulaR1C1 = "Strand"
Range("h1").Select
ActiveCell.FormulaR1C1 = "Chr"
Range("i1").Select
ActiveCell.FormulaR1C1 = "Start"
Range("j1").Select
ActiveCell.FormulaR1C1 = "End"
Range("k1").Select
ActiveCell.FormulaR1C1 = "Position"
Range("l1").Select
ActiveCell.FormulaR1C1 = "Sequence"
Range("m1").Select
ActiveCell.FormulaR1C1 = "TM"
Range("n1").Select
ActiveCell.FormulaR1C1 = "Length"
Range("o1").Select
ActiveCell.FormulaR1C1 = "Total Score"
Range("p1").Select
ActiveCell.FormulaR1C1 = "Cross Hyb"
Range("q1").Select
ActiveCell.FormulaR1C1 = "Delta TM"
Range("r1").Select
ActiveCell.FormulaR1C1 = "Folding"
Range("s1").Select

```

```
ActiveCell.FormulaR1C1 = "Position"
Range("t1").Select
ActiveCell.FormulaR1C1 = "Low Complexity"
```

```
Rows("1:1").Select
Selection.Font.Bold = True
Columns("a:V").Select
Columns("a:V").EntireColumn.AutoFit
```

End Sub

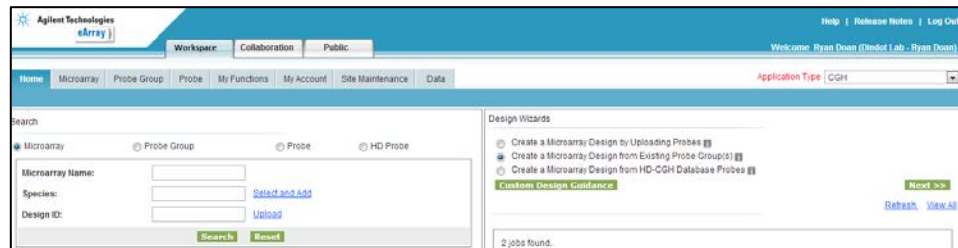
- Once the macro is complete you will see a formatted file (See image below)
 - Save your excel file and proceed to filtering

Transcript ID	Exon #	Range Start	Range End	S	E	Strand	Chr	Start	End	Position	Sequence
ENSBTAT00000034533	0	371876	372126	59	118	-	chr20	372009	372068	chr20:372009-372068	CTCAAGAGTTTG
ENSBTAT00000034533	0	371876	372126	120	177	-	chr20	371950	372007	chr20:371950-372007	CACAAGAATAG
ENSBTAT00000034533	1	368280	368450	40	96	-	chr20	368355	368411	chr20:368355-368411	GACATCCTATGC
ENSBTAT00000034533	1	368280	368450	96	148	-	chr20	368355	368411	chr20:368355-368411	GACATCCTATGC

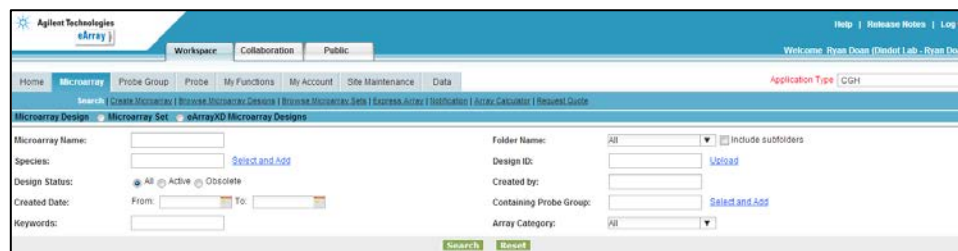
Custom filtering of probes

- Initial filtering by chromosome
 - Removing duplicated regions
 - Select Remove Duplicates in excel
 - Remove duplicates based only on Position (Column K)
 - Select Column for Sequence
 - Select Conditional Formatting (See image below)
 - Select Highlight Cell Rules
 - Select Duplicate Value
 - Popup will appear – any option that fills cells with a color works
 - Click OK

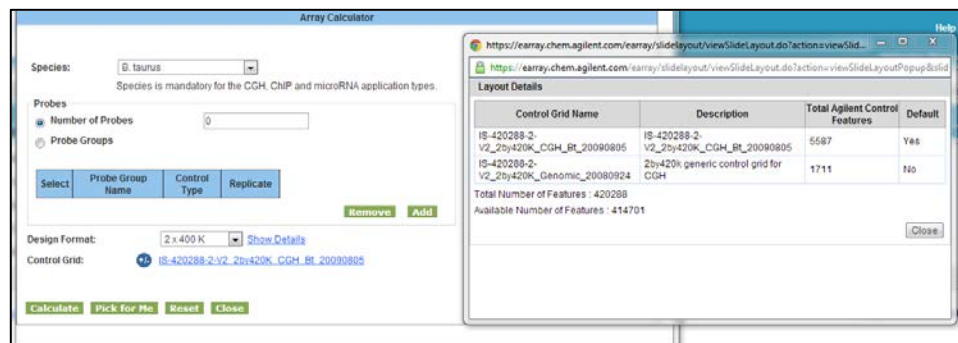
- You should see a Home screen (See image below)
 - Make sure that Application Type is set to CGH



- Select the Microarray Tab (see image below)
 - Click on Array Calculator



- A new window will appear (See image below)
 - Select the species
 - Select the Design format
 - Click on Show Details
 - A new popup will show the number of control probes and available feature on the array.
 - The available features are the total probes you can place on the array



- Write down the number of probes for your design
 - This will be the number of probes to select from your excel sheet

- Close out of all eArray windows
- Final Probe sorting:
 - Open excel file that contains all filtered probes for all chromosomes in a single sheet.
 - Further filter the probes by increasing the cutoffs of the properties
 - TM – remove if >84 or <73
 - Continue filtering until all probes fit on the array design
 - Create a new sheet for the final probe list
 - This will be formatted and uploaded into eArray
 - Formatting to “Complete Array Format”
 - Add Column IDs as follows:
 - GeneSymbol, ID, Location, Sequence, Description, Target ID, Accessions
 - Copy and paste values:
 - Transcript ID from probe list into GeneSymbol column
 - Position into Location column
 - Sequence into Sequence column
 - Create a unique ID for each probe (i.e., chromosome and start position (chr20:34567))
 - Leave remaining columns blank
 - Copy/paste formatted file into text file and save
- Log into eArray
 - Select Probe Tab (See image below)
 - Select Upload
 - Select Species
 - Create probe group – enter ID for group
 - File format = Complete
 - File Type = TDT
 - Select Remove replicate probes from upload
 - Choose your file
 - Click Next

- Select Species
- Click Next

Create a Microarray Design from Existing Probe Group(s)

Select Species	Select Array Type : <input checked="" type="radio"/> Standard CGH <input type="radio"/> CGH+SNP
Define Design	Select Species : <input type="text" value="B. taurus"/>
Layout Probes	
Create Microarray Design	

- A new page will open (See image below)
 - Enter a name for the array
 - Select Design format
 - Check “Append linker to 3’ end”
 - This will make all probes 60bp in length
 - Use the Agilent linker
 - Click next

Create a Microarray Design from Existing Probe Group(s) [Help](#)

Select Species	Design Details	
Define Design	Microarray Name: <input type="text" value="Enter your Array Name here"/>	Species: B. taurus
Layout Probes	Folder: <input type="text" value="Dindot Lab - Ryan Dear"/>	Control Grid: [S-420289-2-1/2_2b/420K_CGH_Bt_20090805]
Create Microarray Design	Design Format: <input type="text" value="2 x 400 K"/> Show Details	Comments: <input type="text"/>
	Description: <input type="text"/>	Attachment: <input type="text"/>
	Keywords: <input type="text"/>	
	Linker Details:	
	<input checked="" type="checkbox"/> Append linker to 3' end	
	Linker Length <input checked="" type="radio"/> Make probes of length <input type="text" value="60"/> <input type="radio"/> Add linker of length <input type="text" value="0"/>	Linker Sequence <input checked="" type="radio"/> Use Agilent linker sequence <input type="text" value="ATAACCGACGCCTAA"/> <input type="radio"/> Use customer linker sequence <input type="text"/>

- A new window will open to select probes for the design (See image below)
 - Click on Add under the Biological CGH Probe Group(s)

Create a Microarray Design from Existing Probe Group(s) [Help](#)

Select Species

Define Design

Layout Probes

Create Microarray Design

Microarray Statistics:

Number of Microarrays per Slide: 2	Percentage Filled by Selected Probe Groups: 1.33%
Number of Slides: 1	Number of Available Features: 414701
Total Number of Features: 420288	Number of User Controls: 0
Number of Agilent Controls: 5587	Total Percentage Filled: 1.33 %
Number of Features Occupied by Replicate Probes: 0	Number of Features Occupied by SNP Probes: 0
Number of Features Occupied by Normalization Probes: 0	

Biological CGH Probe Group(s) Number of Probe Group(s) Selected: 0

Click Add to select probe group(s).

Select	Probe Group Name	Control Type	Replicate
0	Normalization Probe Group Details : Number of Normalization Probe Group(s) Selected: 0		
0	Replicate Probe Group Details : Number of Replicate Probe Group(s) Selected: 0		

Fill Microarrays Feature Layout Randomized Customer Specified

Enable Microarray Set

- A new window will open to locate probe groups
 - In the search box type your group name or leave blank
 - Click search
 - Locate and select your group
 - Click Add to place it into the box to the right
 - Click Done
 - Click Next

<https://earray.chem.agilent.com/earray/common/selectAndAddProbeGroup.do?subAction=search> [Help](#)

Probe Group Category List

Probe Group Name:

8 probe groups found.

Folder:

Probe Group Name ▲	Folder	Status
B Taurus MHC Exons 7-30-09	Dindot Lab - Ryan Doan	Locked
Bovine 180K Exon 9-3-09	Dindot Lab - Ryan Doan	Locked
Bovine Whole Genome Exon 400K 7-31	Dindot Lab - Ryan Doan	Locked
Canine Ensembl64 Exons	Dindot Lab - Ryan Doan	Locked
Canine Exon Array 361438 Probes 1-25-10	Dindot Lab - Ryan Doan	Incomplete
Canine Exon Array 17900 Genes 1-25-10 with 414949	Dindot Lab - Ryan Doan	Incomplete
Horse Whole Genome Exon Array V1 2-16-2010	Dindot Lab - Ryan Doan	Locked
Rat Ensembl59 Autosomal Exons 8-10-2010	Dindot Lab - Ryan Doan	Locked

Selected Probe Group

Bovine Whole Genome Exon 400K 7-31

- A window will open to select the status
 - If the array is complete – select Submitted

- Select Design Checklist
 - A window will open, check all of the items, click ok
- Make sure the checkbox beside Design Checklist is now checked
- Submit array
- You will get an email when it is completed
- Once the design is ready, you can use the AMID number to get quotes or order the array

Create a Microarray Design from Existing Probe Group(s) Help	
Select Species	Create microarray design with this status <input type="radio"/> Draft Allows only you to edit the design. Later, you can change the status to any of the others. <input type="radio"/> Review Lets you and the other users in your workgroup make changes to the design and save new versions of it. <input type="radio"/> Complete Prevents further edits to the design. You must submit the design to Agilent Manufacturing before you can request a quote. <input checked="" type="radio"/> Submitted Submits the design to Agilent Manufacturing, and lets you request a quote for the design. The design cannot be edited. <input type="checkbox"/> Design Checklist
Define Design	
Layout Probes	
Create Microarray Design	

End of Array Design!

APPENDIX 2.3

Array CGH Protocol

Shearing genomic DNA:

- Dilute genomic DNA to 7 µg in 120 µl of elution buffer (Qiagen or Invitrogen)
 - You can use more or less (at least 5 µg), too much DNA will alter the shearing
- Shear DNA using Sonic Dismembrator 500 keeping samples on ice. Adjust setting to reach desired length. Start with 3 – 15 sec pulses at 12% power with a 30 sec break between pulses.
- Run 10 µl of sample on gel to confirm size (should see a smear of DNA with the majority at the size desired. Once a setting is chosen, this does not need to be completed for every sample
- Purify DNA using the Invitrogen Purelink PCR Kit and elute in 30 µl EB (perform 2 washes (600 µl followed by 300 µl))
- Determine quantity of DNA using NanoDrop

Cy3 and Cy5 Labeling:

- Add 4 µg of sheared/purified DNA in 24 µl to PCR tube and add 20 µl cy3 (Alexa Fluor 555) or cy5 (Alexa Fluor 647) labeled panomers to the tubes. Use 555 for control and 647 for sample.
- Mix and spin down briefly.
- Place in thermocycler for 10 minutes at 95°C.
- Place samples on ice for 5 minutes
- Add 5 µl 10X nucleotide mix with the same dye label as the panomers added.
- Mix and spin down briefly.
- Place in thermocycler at 37°C overnight.
- Purify with Invitrogen Purelink PCR Kit and elute in 30 µl EB (perform 2 washes (600 µl followed by 300 µl))
- Quantitate DNA and labeling efficiency using the microarray function on the NanoDrop

- The DNA should be several hundred nanograms/ μl and the labeling should be at least 5 pmol/ μl

Hybridization: Follow steps below for specific array format.

8x60K Array:

- Mix 2 μg of control labeled DNA (555) with 2 μg of sample (647) with a total volume of 16 μl
- Add to each reaction:
 - 2 μl Cot-1 DNA
 - 4.5 μl Agilent 10X Blocking Buffer
 - 22.5 μl 2X HiRPM Buffer
- Mix gently, trying to avoid bubbles
- Spin down briefly
- Place on heat block at 95°C for 3 minutes
- Place in 37°C hot water bath for 30 minutes
- Briefly spin down
- Place gasket slide in hybridization clamp with rubber side up
- Slowly add 40 μl of sample to array gasket slide (avoid liquid contact with rubber gasket)

2x400 thousand array:

- Mix 2 μg of control labeled DNA (555) with 2 μg of sample (647) with a total volume of 79 μl
- Add to each reaction:
 - 25 μl Cot-1 DNA
 - 26 μl Agilent 10X Blocking Buffer
 - 130 μl 2X HiRPM Buffer
- Mix gently, trying to avoid bubbles
- Spin down briefly
- Place on heat block at 95°C for 3 minutes
- Place in 37°C hot water bath for 30 minutes

- Briefly spin down
- Place gasket slide in hybridization clamp with rubber side up
- Slowly add 240 μ l of sample to array gasket slide (avoid liquid contact with rubber gasket)

1x1 million array:

- Mix 4 μ g of control labeled DNA (555) with 4 μ g of sample (647) with a total volume of 158 μ l
 - Add to each reaction:
 - 50 μ l Cot-1 DNA
 - 52 μ l Agilent 10X Blocking Buffer
 - 260 μ l 2X HiRPM Buffer
 - Mix gently, trying to avoid bubbles
 - Spin down briefly
 - Place on heat block at 95°C for 3 minutes
 - Place in 37°C hot water bath for 30 minutes
 - Briefly spin down
 - Place gasket slide in hybridization clamp with rubber side up
 - Slowly add 490 μ l of sample to array gasket slide (avoid liquid contact with rubber gasket)
-
- Apply array slide with side labeled “Agilent” down to the gasket.
 - Place top of clamp on slide and tighten.
 - Slowly turn slide/clamp unit to ensure no bubbles are stuck in place, if so, tap them loose
 - Place in 65°C hybridization oven for 24 hours with a rotational speed of 20 RPM
 - Place 500 mL of Wash Buffer 2 in bottle a warm overnight in 37°C water bath.

Washing array

- Add 400 mL of Wash Buffer 1 into two clean glass dishes.
- Remove clamp from slide and gasket

- Submerge array slide with array gasket in wash buffer 1 and slide plastic forceps between the two slides with gasket on the bottom, twist and allow gasket to fall to bottom of dish
- Place slide in slide rack submersed in 2nd dish of wash buffer 1 for 5 minutes
- Pour wash buffer 2 into new glass dish
- Quickly transfer slide and rack to wash buffer 2, minimizing air contact – 5 minutes with gentle agitation for 30 seconds, every 30 seconds
- Remove and place in 4th glass dish with acetonitrile 30 sec
- Slowly remove from acetonitrile, avoiding any bubbles or spots on the glass
- Place slide in slide holder with Agilent label facing up
- Place ozone barrier on top of slide and close holder
- Scan slides using the following setting:
 - 2 μ m resolution
 - 0.05 XDR
- 2 Color CGH
- Export data from array images using Agilent Feature Extraction software

Reagent Used:

1. Invitrogen Purelink PCR Purification Kit (50 rxns)
\$92.00

The PCR purification system is used for both post-sonication and DNA labeling purifications.

2. Invitrogen Bioprime Plus CGH Genomic Labeling Module (30 rxns)
\$752.00

The genomic labeling kit from Invitrogen is required to label the DNA samples prior to hybridization on the arrays. Each array will require two labeling reactions, cy3 and cy5. The control sample will be labeled the same color to allow multiple arrays to be compared against each other.

3. Oligo aCGH/Chip-Chip Hybridization Kit (Agilent) \$361.50

The hybridization kit provides the necessary solutions to ensure that each sample is able to bind to the array. The kit provides enough reagent to perform hybridizations on 25 slides.

4. Agilent SurePrint G3 Custom CGH Microarray 1 x 1M \$535.50

4.1 Agilent SurePrint G3 Custom CGH Microarray 2 x 400K ~\$650*

These slides will be used a total of 2 times to allow for 2 experiments from a single array.

5. 1 x 1M Backings (pack of 5) (or 2x400k) \$114.00

The backings are required to seal the arrays during hybridizations. Each backing will be used 1 time. Therefore, the total of 1 slide hybridizations will require 2 packs of backings.

6. Oligo aCGH/ChIP-on-chip Wash Buffer Kit \$246.75

The wash buffer kit is required to clean the array slides following hybridization. A single kit will wash 40 array slides.

7. Cot-1 Human DNA (Invitrogen) \$180.90

The Cot-1 is used during hybridization to reduce noise on the array from repetitive DNA

8. Ozone-Barrier Slide Cover Kit (Agilent) 25 covers \$503.00*

Ozone barrier slides are placed over the array during the scanning process to prevent degradation of the array by ozone during scanning.

APPENDIX 2.4

CNV PCR PRIMERS

Index	Forward Primer	Reverse Primer	Product Length	Gene ID
1	CAGCATTATCAAGCCAGGT	GGAGGGTGGTAGTGGTGTTT	217 bp	ENSBTAG00000022376
2	GAATACCCCATGCTTCAGA	CCACCTGGACACTGGTTAGC	212 bp	ENSBTAG00000033381
3	CCCTGATTTTGGATGTCTGG	CCCACACCAATCTGACCAT	289 bp	ENSBTAG00000000120
4	CCCCTTCAGGTTACTCCACA	CCTGCTTCCTCTCCTCCTTC	205 bp	ENSBTAG00000019524
5	GTCTGACCTCCAGGCTCCTC	CCTCAGAATGGGCCAGATAG	218 bp	ENSBTAG00000006630
6	GCTTCATTTGGGAGCTGAG	CCTCTGTGCGAGACCTTCA	205 bp	ENSBTAG00000021747
7	TCTGCTTTTTGGTTTGGAGCA	GTTGACAAGGCAGCTTCTCC	251 bp	ENSBTAG00000024826
8	AGCAGCAATTTTACCGTGT	CAAGAAACGTGCTGCCTATG	206 bp	ENSBTAG00000017743
9	CCCAAAGGAAGAAGTCGATG	CCATGTCCTTTTTCCCTCT	258 bp	ENSBTAG00000015780
10	GGATCGCACAAACCAATAACC	GCACACATCTGGCTGTTCTG	213 bp	ENSBTAG00000011932
11	TGCTCAGGATGGTGGTGATA	CTTTCTGGACGCACTCATCA	227 bp	ENSBTAG00000016317
12	TTCTGGCAAAGTGGACATCG	GCCTTGACTGTGCCGTTGA	205 bp	ENSBTAG00000030368
13	GCTTTGTGAAGCTCCCTGTC	TCCCATAGGATTCTGGCTTC	415 bp	ENSBTAG00000005742
14	CATTACCCAAATTTGGTGCTTT	CATCAATTTCTCGGTGCTCA	832 bp	ENSBTAG00000033169
15	CCTCAAACCTGAAGCAAAA	GAGGCCCACTGACTACCAGA	711 bp	ENSBTAG00000007644
16	AAGGCTCAAGGAGCTCATCA	TGCACCATGCTCACAAATTC	122 bp	ENSBTAG00000014402
17	AGGCCATTACTGCGAGAAGA	ACAACCACGATCCCAAAGAT	135 bp	ENSBTAG00000023623
18	ATCCTAGCCGCAACACATC	ATCGAGGGATGACTACCAC	130 bp	ENSBTAG00000038245
19	ATTCAGCTGATCTGGCTGGT	CAGGGATGGTTGCTCCTTC	105 bp	ENSBTAG00000043673
20	CACGGTGTTCCTGAAAAGT	GGAACCTCATTGGAGCTGAA	122 bp	ENSBTAG00000005941
21	CCAGCCTCAAGAGAGTCACC	ACTGGCCATGTTGAAGAAC	122 bp	ENSBTAG00000020385
22	CGTCATACGAGGGGTAGTCG	GCCTGTCCGGAGGATATTTT	143 bp	ENSBTAG00000009556
23	GACCTCCTTCCTACCCAACG	GCTCACAAAGAGACCTCGAA	151 bp	ENSBTAG00000031685
24	GCACATCTCCACCTTCATCA	CCTCCTCCACTTTCTTCACG	118 bp	ENSBTAG00000006059
25	ACCATAGAGGAGTGGTGAA	ATGACCCTATGTTGCCATC	130 bp	ENSBTAG00000031097
26	GGAGCCAGCTGTAATGAGG	TGTGGACCAATCGAGTCAA	110 bp	ENSBTAG00000025994
27	GGGGAAGTGTCACAGAGGAA	CTGCCATCTGTAAGCCATTGT	177 bp	ENSBTAG00000031100
28	GGGGCAGTAATCACAGGAGA	CATGGCTAGGGGATTCTT	106 bp	ENSBTAG00000009999
29	GTGCCATGTTAGTGGTCACG	CAAGGAAGAAAGCCACAGA	127 bp	ENSBTAG00000038953
30	TCAAGTAGGTGCAAAGCTCAA	CAGACGTTTGTGGGTTGTG	150 bp	ENSBTAG00000038720
31	TGTGGGGATTTTCCAGTTTC	TGTGAAGGCTGTGATTGAGC	128 bp	ENSBTAG00000037384
32	TTCTCTGAGTAGCCGATGG	TTTGGCACTCTCATGCTCAG	152 bp	ENSBTAG00000015219
33	ATGGCCAAAACACTCAAAGG	CATCGGTATGGCTCTCCAGT	122 bp	ENSBTAG00000009478
34	GCTTCAGCAAGGACCAAGTC	TCCTTTGGTTTTGCTGCTTT	132 bp	ENSBTAG00000009478
35	TTCTGGCAAAGTGGACATCG	GCCTTGACTGTGCCGTTGA	205 bp	GAPDH
36	CTGGGAAACCAAACCTTGCAT	TTGCAGCAGCATGGATAGAC	145 bp	Intergenic
37	TGGGCCACCAAGTAACTTAG	TCACGTATGGCAAAAACCAA	107 bp	Intergenic
38	ATGTATGCCAACCTTCAGC	GCTTCACTTGCTGGTTCTC	123 bp	Intergenic
39	AGCATAAGGGGGGAGAAAAAT	GCCCCGCTATTATAGAACCA	101 bp	Intergenic
40	GCTACTGAGCCACTGGGAAC	GCTACCTGCTCCTTGCTGG	99 bp	Intergenic
41	CCTGCAGAAGGAAATGGAAA	CACTGTGCTGTGGAAGTGGT	139 bp	Intergenic
42	CCATCACCCACAACCTCCTCT	CCTTGCTTTGGCTTTTTGAG	134 bp	Intergenic
43	TGGGGTGACTACCTTTCTGC	GGATAGCTGGTCCCTCAACA	142 bp	Intergenic
44	GCCATAAGGGTGGTGAAGAA	GCCTTCTCCCAAATGCTGTA	123 bp	Intergenic
45	TCAGCCCTCTCCTACCTCAA	GGAGGATTTCCCACTGGATT	112 bp	Intergenic
46	TAAGGCATTGCATCAACCAA	TGAGCAACAACAACACAGCA	123 bp	Intergenic
47	ACGGGACTTCCCATATTTCC	GTTTTGGATTTTGGCAGGA	120 bp	Intergenic
48	CAACATGGGTGCAACAGAAG	GGTGGCTGTGCTCAATTTCT	120 bp	Intergenic
49	GCAGGAAGGTAGTTGGGTCA	GTTGTTGGATTGCTGGCTTT	116 bp	Intergenic
50	CGCAGTGGGATGGAAGTATT	GCTATGATCCGCTCTGGAAC	102 bp	Intergenic
51	TGCCACAAGAGAGGTCACTG	ATTTTGGCTGTGCTGAATCC	113 bp	Intergenic
52	GAGCACACGGTCTCCTTAC	GGAGCAGAGACACAGAGCAGA	130 bp	Cath1
53	CCATTTCCAGGTTAGGATGAC	GAGGTGTGGTGTGTGTGG	114 bp	Cath4

APPENDIX 2.5
QPCR PROTOCOL

SYBR Green Protocol:

SYBR GreenER qPCR SuperMix for ABI PRISM = 11760-500

Reagent	Volume (1 Reaction)
SYBR GreenER	5 μ l
Forward Primer	0.2 μ l
Reverse Primer	0.2 μ l
Water	3.1 μ l
DNA (25 ng)	1.5 μ l

- Run all samples in triplicate
- Data analysis: $\Delta\Delta$ CT Method

APPENDIX 2.6

LOW CONFIDENCE CNVS

	n	Copy Number Variants (CNVs)	CNV Regions (CNVRs)	CNVs / Animal	CNVR Gains	CNVR Losses	Complex CNVRs	Variant Genes	Deletion Variant Genes
Total	27	363	162	13	53	109	0	213	8

APPENDIX 2.7

CNV GENES

Please see attached Microsoft Excel file for appendix 2.7 containing a list of all genes affected by CNVs from CGH CNV analyses. The table contains columns for: ensembl gene ID, chromosome, start, end, status, and biotype.

APPENDIX 2.8

CNV CGH BP STATISTICS

Exome CNV Gene BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	96	52.58070803	1.30E-08
Signal Transduction	135	22.90512888	1.70E-06
Sensory Perception	127	32.6157858	8.27E-08
Regulation of Cell Cycle	27	7.355582061	0.061387616
Cellular Process	13	4.518775466	0.104414395
Metabolism	8	3.475547636	0.062282103
Protein processing	-	-	-
Nucleic Acid Metabolism	-	-	-
Miscellaneous	-	-	-
Developmental Process	-	-	-
Exome Deletion Gene BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	12	19.00214713	0.00027312
Signal Transduction	7	3.166895484	0.075145159
Sensory Perception	7	4.078825867	0.043423823
Cellular Process	-	-	-
Metabolism	-	-	-
Nucleic Acid Metabolism	-	-	-
Protein processing	-	-	-
Miscellaneous	-	-	-
Regulation of Cell Cycle	-	-	-
Developmental Process	-	-	-
Exome Taurus CNV Gene BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	98	68.28944398	9.4744E-11
Signal Transduction	83	16.0209536	6.26453E-05
Sensory Perception	80	23.52345158	7.79736E-06
Miscellaneous	-	-	-
Metabolism	-	-	-
Protein processing	3	2.578468508	0.108326425
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	-	-	-
Cellular Process	-	-	-
Exome Indicus CNV BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	62	39.49631401	5.72E-07
Signal Transduction	112	29.85900834	4.65E-08
Sensory Perception	109	39.95211377	1.09E-08
Miscellaneous	-	-	-
Metabolism	-	-	-
Protein processing	3	2.312912233	0.128303466
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	52	16.71794767	0.010377683
Cellular Process	16	7.407706929	0.059977884

Appendix 2.8 Continued

Exome Angus CNV BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	78	63.29236914	1.05E-10
Signal Transduction	75	15.30688248	9.14E-05
Sensory Perception	72	21.97821752	1.69E-05
Miscellaneous	-	-	-
Metabolism	-	-	-
Protein processing	3	2.768717952	0.096123121
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	-	-	-
Cellular Process	-	-	-
Exome Holstein CNV BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	26	25.23259145	1.38E-05
Signal Transduction	37	13.89559858	0.00019325
Sensory Perception	35	16.62028393	4.57E-05
Miscellaneous	-	-	-
Metabolism	-	-	-
Protein processing	-	-	-
Developmental Process	-	-	-
Nucleic Acid Metabolism	3	2.274520494	0.131515795
Regulation of Cell Cycle	-	-	-
Cellular Process	-	-	-
Exome Nellore CNV BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	69	48.7376226	7.14E-08
Signal Transduction	92	23.33079185	8.59E-06
Sensory Perception	88	29.64924845	1.64E-06
Miscellaneous	3	2.032164743	0.15400112
Metabolism	5	2.222841578	0.135982576
Protein processing	3	2.572738694	0.108719365
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	12	4.021338627	0.133899024
Cellular Process	3	2.328856145	0.126995166
Exome Brahman CNV BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	4	2.644934083	0.103880466
Signal Transduction	64	33.79041152	6.14E-09
Sensory Perception	63	42.34781581	7.64E-11
Miscellaneous	-	-	-
Metabolism	-	-	-
Protein processing	-	-	-
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	9	6.441646575	0.091992446
Cellular Process	-	-	-
Fisher Exact Test 2-sided p value Exome CNVs			
Term	Taurine Genes	Indicine Genes	p value
Immunity and Defense	98	62	3.92E-05
Signal Transduction	83	112	1
Sensory Perception	80	109	0.933306
Miscellaneous	-	-	-
Metabolism	-	-	-
Protein processing	3	3	0.704745
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	0	52	1.60E-12
Cellular Process	0	16	0.000173103

Appendix 2.8 Continued

Fisher Exact Test 2-sided p value Exome CNVs			
Term	Angus Genes	Nellore Genes	p value
Immunity and Defense	78	69	0.110415
Signal Transduction	75	92	1
Sensory Perception	72	88	1
Miscellaneous	0	3	0.255853
Metabolism	0	5	0.0678283
Protein processing	3	3	1
Developmental Process	-	-	-
Nucleic Acid Metabolism	-	-	-
Regulation of Cell Cycle	0	12	0.00161931
Cellular Process	0	3	0.255853
All CNV Genes BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	190	34.11575456	3.87E-05
Signal Transduction	293	22.63587265	1.22E-05
Sensory Perception	250	25.09718707	5.45E-07
Regulation of Cell Cycle	-	-	-
Cellular Process	8	2.08394325	0.148855201
Metabolism	34	7.726892954	0.052006439
Protein processing	4	2.389751815	0.122133017
Nucleic Acid Metabolism	-	-	-
Miscellaneous	68	7.456127419	0.058696507
Developmental Process	17	4.099309178	0.128779378
All Taurus CNV Genes BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	170	53.19609523	6.85E-08
Signal Transduction	221	28.63521331	8.74E-08
Sensory Perception	202	39.29604703	2.93E-09
Regulation of Cell Cycle	-	-	-
Cellular Process	-	-	-
Metabolism	-	-	-
Protein processing	-	-	-
Nucleic Acid Metabolism	-	-	-
Miscellaneous	-	-	-
Developmental Process	-	-	-
All Indicus CNV Genes BP			
Term	Genes	Fishers Method	p value
Immunity and Defense	147	30.36319814	8.14E-05
Signal Transduction	223	17.44666697	0.000162744
Sensory Perception	193	20.29308567	3.92E-05
Regulation of Cell Cycle	18	2.107396493	0.146588527
Cellular Process	5	2.497279133	0.114043174
Metabolism	28	7.26031032	0.064047732
Protein processing	-	-	-
Nucleic Acid Metabolism	-	-	-
Miscellaneous	24	7.168947686	0.06670341
Developmental Process	30	7.527378776	0.056859132

APPENDIX 2.9

CNV UMD3.1 CNV STATISTICS

CNV Gene BP		
Term	Genes	p value
Immunity and Defense	116	1.69381E-09
Signal Transduction	98	0.000802219
Sensory Perception	97	0.000125488
Regulation of Cell Cycle	0	-
Cellular Process	9	0.095763784
Metabolism	6	0.130279399
Protein processing	11	0.113480573
Nucleic Acid Metabolism	6	0.130279399
Miscellaneous	0	-
Developmental Process	0	-

APPENDIX 2.10

CNV GENES WITH ASSOCIATED OMIM TERMS

OMIM # Morbidity Map	Gene ID	Phenotype
131244	<i>EDNRB</i>	ABCD syndrome,
605080	<i>CNGB3</i>	Achromatopsia-3,
607008	<i>ACADM</i>	Acyl-CoA dehydrogenase, medium chain, deficiency of,
610613	<i>CYP11B1</i>	Adrenal hyperplasia, congenital, due to 11-beta-hydroxylase deficiency,
146770	<i>IGLL1</i>	Agammaglobulinemia 2,
601920	<i>JAG1</i>	Alagille syndrome,
100650	<i>ALDH2</i>	Alcohol sensitivity, acute,
147450	<i>SOD1</i>	Amyotrophic lateral sclerosis, due to SOD1 deficiency,
607465	<i>CDAN1</i>	Anemia, congenital dyserythropoietic, type I,
602322	<i>TERC</i>	Aplastic anemia,
605010	<i>SPINK5</i>	Atopy,
602617	<i>TTF2</i>	Bamforth-Lazarus syndrome,
170261	<i>TAP2</i>	Bare lymphocyte syndrome, type I, due to TAP2 deficiency,
134370	<i>CFH</i>	Basal laminar drusen,
120250	<i>COL6A3</i>	Bethlem myopathy,
608614	<i>CYP4V2</i>	Bietti crystalline corneoretinal dystrophy,
603248	<i>BMPR1B</i>	Brachydactyly, type A2,
190080	<i>MYC</i>	Burkitt lymphoma,
607844	<i>LEMD3</i>	Buschke-Ollendorff syndrome,
604283	<i>PRG4</i>	Camptodactyly-arthropathy-coxa vara-pericarditis syndrome,
176872	<i>MAP2K1</i>	Cardiofaciocutaneous syndrome,
610897	<i>CHMP4B</i>	Cataract, posterior polar, 3,
610933	<i>LRSAM1</i>	Charcot-Marie-Tooth disease, axonal, type 2P,
600635	<i>TTF1</i>	Chorea, hereditary benign,
608512	<i>NCF1</i>	Chronic granulomatous disease due to deficiency of NCF-1,
600678	<i>MSH6</i>	Colorectal cancer, hereditary nonpolyposis, type 5,
614123	<i>TMCO1</i>	Craniofacial dysmorphism, skeletal anomalies, and mental retardation syndrome
191740	<i>UGT1A1</i>	Crigler-Najjar syndrome, type I,
130160	<i>ELN</i>	Cutis laxa, AD,
611716	<i>ATP6V0A2</i>	Cutis laxa, autosomal recessive, type IIA,
607657	<i>CTH</i>	Cystathioninuria,
604175	<i>RPL11</i>	Diamond-Blackfan anemia 7,

Appendix 2.10 Continued

OMIM # Morbidity Map	Gene ID	Phenotype
605849	<i>DMGDH</i>	Dimethylglycine dehydrogenase deficiency,
602880	<i>GDF1</i>	Double-outlet right ventricle,
182860	<i>SPTA1</i>	Elliptocytosis-2,
107400	<i>SERPINA1</i>	Emphysema due to AAT deficiency,
602208	<i>KCNJ10</i>	Enlarged vestibular aqueduct, digenic,
607566	<i>EPM2A</i>	Epilepsy, progressive myoclonic 2A,
608072	<i>EPM2A</i>	Epilepsy, progressive myoclonic 2B,
602926	<i>STXBP1</i>	Epileptic encephalopathy, early infantile, 4,
601011	<i>CACNA1A</i>	Episodic ataxia, type 2,
604579	<i>FZD4</i>	Exudative vitreoretinopathy,
612309	<i>F5</i>	Factor V deficiency,
613899	<i>FANCC</i>	Fanconi anemia, complementation group C,
138160	<i>SLC2A2</i>	Fanconi-Bickel syndrome,
600968	<i>SLC12A3</i>	Gitelman syndrome,
238300	<i>GLDC</i>	Glycine encephalopathy,
610860	<i>AGL</i>	Glycogen storage disease IIIa,
150000	<i>LDHA</i>	Glycogen storage disease XI,
603868	<i>RAB27A</i>	GrisCELLI syndrome, type 2,
602365	<i>CTSC</i>	Haim-Munk syndrome,
606857	<i>GCLC</i>	Hemolytic anemia due to gamma-glutamylcysteine synthetase deficiency,
603401	<i>AP3B1</i>	Hermansky-Pudlak syndrome 2,
607521	<i>HPS5</i>	Hermansky-Pudlak syndrome 5,
138130	<i>GLUD1</i>	Hyperinsulinism-hyperammonemia syndrome,
601199	<i>CASR</i>	Hyperparathyroidism, neonatal,
109700	<i>B2M</i>	Hypoproteinemia, hypercatabolic,
164050	<i>PNP</i>	Immunodeficiency due to purine nucleoside phosphorylase deficiency,
147200	<i>IGKC</i>	Kappa light chain deficiency,
606890	<i>GALC</i>	Krabbe disease,
116897	<i>CEBPA</i>	Leukemia, acute myeloid,
606686	<i>EIF2B1</i>	Leukoencephalopathy with vanishing white matter,
611966	<i>TRAPPC9</i>	Mental retardation, autosomal recessive 13,
300646	<i>ZDHHC9</i>	Mental retardation, X-linked syndromic, Raymond type,
605452	<i>ABCB6</i>	Microphthalmia, isolated, with coloboma 7,
602153	<i>KRT81</i>	Monilethrix,
602765	<i>KRT83</i>	Monilethrix,
601928	<i>KRT86</i>	Monilethrix,

Appendix 2.10 Continued

OMIM # Morbidity Map	Gene ID	Phenotype
605073	<i>TRIM37</i>	Mulibrey nanism,
607939	<i>SUMF1</i>	Multiple sulfatase deficiency,
156225	<i>LAMA2</i>	Muscular dystrophy, congenital merosin-deficient,
102565	<i>FLNC</i>	Myopathy, distal, 4,
607215	<i>NPHP4</i>	Nephronophthisis 4,
602716	<i>NPHS1</i>	Nephrotic syndrome, type 1,
614297	<i>C19orf12</i>	Neurodegeneration with brain iron accumulation 4,
608581	<i>RP1L1</i>	Occult macular dystrophy,
148042	<i>KRT6B</i>	Pachyonychia congenita, Jackson-Lawler type,
148041	<i>KRT6A</i>	Pachyonychia congenita, Jadassohn-Lewandowsky type,
601501	<i>VPS35</i>	Parkinson disease 17,
609007	<i>LRRK2</i>	Parkinson disease-8,
609023	<i>PNKD</i>	Paroxysmal nonkinesigenic dyskinesia,
607751	<i>TAS2R38</i>	Phenylthiocarbamide tasting,
603390	<i>PDE8B</i>	Pigmented nodular adrenocortical disease, primary, 3,
600565	<i>NRXN1</i>	Pitt-Hopkins-like syndrome 2,
606938	<i>UROS</i>	Porphyria, congenital erythropoietic,
314310	<i>TFE3</i>	Renal cell carcinoma, papillary, 1,
603345	<i>SLC4A5</i>	Renal tubular acidosis, proximal, with ocular abnormalities,
613596	<i>FAM161A</i>	Retinitis pigmentosa 28,
609507	<i>TOPORS</i>	Retinitis pigmentosa 31,
608400	<i>USH2A</i>	Retinitis pigmentosa 39,
600342	<i>RGR</i>	Retinitis pigmentosa 44,
180430	<i>RPIA</i>	Ribose 5-phosphate isomerase deficiency,
300642	<i>SRPX2</i>	Rolandic epilepsy, mental retardation, and speech dyspraxia,
606000	<i>BTNL2</i>	Sarcoidosis, susceptibility to,
603005	<i>PAPSS2</i>	SEMD, Pakistani type,
605837	<i>HERC2</i>	Skin/hair/eye pigmentation 1, blond/brown hair,
607642	<i>RAI1</i>	Smith-Magenis syndrome,
610844	<i>SPG11</i>	Spastic paraplegia-11,
611605	<i>ERLIN2</i>	Spastic paraplegia-18,
612641	<i>ANK1</i>	Spherocytosis, type 1,
614154	<i>NOP56</i>	Spinocerebellar ataxia 36,
151443	<i>LIFR</i>	Stuve-Wiedemann syndrome/Schwartz-Jampel type 2 syndrome,
601284	<i>ACVRL1</i>	Telangiectasia, hereditary hemorrhagic, type 2,
606370	<i>TPK1</i>	Thiamine metabolism dysfunction syndrome 5,

Appendix 2.10 Continued

OMIM # Morbidity Map	Gene ID	Phenotype
612025	<i>IYD</i>	Thyroid dysmorphogenesis 4,
613715	<i>POLR1D</i>	Treacher Collins syndrome 2,
600221	<i>TEK</i>	Venous malformations, multiple cutaneous and mucosal,
609506	<i>CYP27B1</i>	Vitamin D-dependent rickets, type I,
602357	<i>WIPF1</i>	Wiskott-Aldrich syndrome 2,
611507	<i>CISD2</i>	Wolfram syndrome 2,
601593	<i>BARD1</i>	Breast cancer, susceptibility to
614295	<i>BICC1</i>	Renal dysplasia, cystic, susceptibility to
602452	<i>BUB1</i>	Colorectal cancer with chromosomal instability
134371	<i>CFHR1</i>	Hemolytic uremic syndrome, atypical, susceptibility to
605336	<i>CFHR3</i>	Hemolytic uremic syndrome, atypical, susceptibility to
604332	<i>CHIC2</i>	Leukemia, acute myeloid
609512	<i>CHMP2B</i>	Amyotrophic lateral sclerosis, CHMP2B-related
118503	<i>CHRNA3</i>	Lung cancer susceptibility 2
124080	<i>CYP11B2</i>	Aldosterone to renin ratio raised
124030	<i>CYP2D6</i>	Codeine sensitivity
606518	<i>HAVCR1</i>	Atopy, resistance to
146880	<i>HLA-DQA1</i>	Celiac disease, susceptibility to
604305	<i>HLA-DQB1</i>	Celiac disease, susceptibility to
142857	<i>HLA-DRB1</i>	Multiple sclerosis, susceptibility to
602376	<i>IFNAR2</i>	Hepatitis B virus, susceptibility to
147620	<i>IL6</i>	Crohn disease-associated growth failure
609269	<i>KIAA0319</i>	Dyslexia, susceptibility to, 2
150270	<i>LAP</i>	Laryngeal adductor paralysis
603025	<i>LAP</i>	Leukemia, acute T-cell lymphoblastic
153245	<i>LEF1</i>	Sebaceous tumors, somatic
108962	<i>NPR3</i>	Hypertension, salt-resistant
164350	<i>OAS1</i>	Diabetes mellitus, type 1, susceptibility to
600632	<i>OPCML</i>	Ovarian cancer, somatic
168820	<i>PON1</i>	Coronary artery disease, susceptibility to
107280	<i>SERPINA3</i>	Alpha-1-antichymotrypsin deficiency
603028	<i>TLR2</i>	Colorectal cancer, susceptibility to
191342	<i>UCHL1</i>	Parkinson disease 5, susceptibility to

APPENDIX 2.11

CNV GENES WITH FST GREATER THAN 0.25

Gene	Total	Angus	Holstein	Brahman	Nellore	Taurus	Indicus	CNV Location	Breeds Fst	Subspecies Fst (Pop)
<i>PSMB7</i>	4	0	0	0	4	0	4	Muliple Exons	0.6829	0.4055
<i>FAM53C</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>ACVRL1</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>TAX1BP1</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>ASZ1</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>CDC27</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>UFM1</i>	4	0	0	0	4	0	4	Muliple Exons	0.5766	0.3407
<i>CERS5</i>	4	0	0	0	4	0	4	Muliple Exons	0.5766	0.3407
<i>PPP1R14C</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>ACADM</i>	4	0	0	0	4	0	4	Muliple Exons	0.5766	0.3407
<i>ARF1</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>CEBPG</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>IFT27</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>LYPLAL1</i>	4	0	0	0	4	0	4	Muliple Exons	0.5766	0.3407
<i>PNKD</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>PON1</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>SRPX2</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>TM7SF3</i>	4	0	0	0	4	0	4	Exon	0.5766	0.3407
<i>TMBIM1</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>TNMD</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>UNC5D</i>	4	0	0	0	4	0	4	Genes	0.5766	0.3407
<i>FANCC</i>	14	12	0	2	0	12	2	Genes	0.5371	0.1794
<i>IGLL1</i>	14	12	0	2	0	12	2	Genes	0.5371	0.1794
<i>TSPY</i>	14	12	0	2	0	12	2	Genes	0.5371	0.1794
<i>ZNF280A</i>	14	12	0	2	0	12	2	Genes	0.5371	0.1794
<i>ZNF280B</i>	14	12	0	2	0	12	2	Genes	0.5371	0.1794
<i>MARCKSL1</i>	5	0	0	1	4	0	5	Exon	0.4673	0.4203
<i>YWHAZ</i>	5	0	0	1	4	0	5	Genes	0.4673	0.4203
<i>KIF20B</i>	5	0	0	1	4	0	5	Genes	0.4673	0.4203
<i>KLRF1</i>	5	0	0	4	1	0	5	Genes	0.4673	0.4203
<i>DMGDH</i>	5	1	0	0	4	1	4	Exon	0.4419	0.2282
<i>NUDT7</i>	5	1	0	0	4	1	4	Exon	0.4419	0.2282
<i>BIRC5</i>	6	0	0	2	4	0	6	Exon	0.4407	0.4944
<i>GOLPH3L</i>	6	0	0	2	4	0	6	Genes	0.4407	0.4944
<i>SPINK5</i>	3	0	0	0	3	0	3	Muliple Exons	0.4284	0.2548
<i>CLK2</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>NES</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>SMG5</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>TSN</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>CSNK2B</i>	3	0	0	0	3	0	3	Muliple Exons	0.4284	0.2548
<i>ANAPC13</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>FLNC</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>ARMC2</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>BARD1</i>	3	0	0	0	3	0	3	Muliple Exons	0.4284	0.2548
<i>BMPR1B</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>C14H8orf47</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>C8H9orf85</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>CCR9</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>CLDND1</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>CNN2</i>	3	0	0	0	3	0	3	Genes	0.4284	0.2548
<i>CTH</i>	3	0	0	0	3	0	3	Muliple Exons	0.4284	0.2548
<i>EEF2K</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548
<i>FOLH1</i>	3	0	0	0	3	0	3	Exon	0.4284	0.2548

FRMD5	3	0	0	0	3	0	3	Exon	0.4284	0.2548
GNG7	3	0	0	0	3	0	3	Exon	0.4284	0.2548
HNRNPM	3	0	0	0	3	0	3	Multiple Exons	0.4284	0.2548
JOSD1	3	0	0	0	3	0	3	Exon	0.4284	0.2548
KCNAB1	3	0	0	0	3	0	3	Genes	0.4284	0.2548
PCMTD1	3	0	0	0	3	0	3	Exon	0.4284	0.2548
SET	3	0	0	0	3	0	3	Multiple Exons	0.4284	0.2548
SOD1	3	0	0	0	3	0	3	Genes	0.4284	0.2548
UGT2B10	3	0	0	0	3	0	3	Genes	0.4284	0.2548
VTA1	3	0	0	0	3	0	3	Multiple Exons	0.4284	0.2548
ZC3H11A	3	0	0	0	3	0	3	Exon	0.4284	0.2548
ZNF804B	4	0	0	0	4	0	4	Multiple Exons	0.4284	0.2548
CFH	3	0	3	0	0	3	0	Genes	0.4284	0.0162
LOC790886	3	0	3	0	0	3	0	Genes	0.4284	0.0162
BLA-DQB	7	3	0	0	4	3	4	Genes	0.4039	0.1712
SIRPB1	6	0	0	3	3	0	6	Genes	0.3797	0.4944
ABCF2	6	0	1	1	4	1	5	Multiple Exons	0.3797	0.3145
ATP6V1E1	7	0	2	1	4	2	5	Multiple Exons	0.3615	0.2312
NXF3	6	1	1	0	4	2	4	Multiple Exons	0.3565	0.1471
CHORDC1	6	2	0	0	4	2	4	Exon	0.3501	0.1471
AOX1	7	1	0	2	4	1	6	Multiple Exons	0.341	0.3963
LOC618367	3	0	0	3	0	0	3	Multiple Exons	0.3221	0.1938
SERPINE2	5	0	0	3	2	0	5	Exon	0.322	0.4203
SON	5	0	0	2	3	0	5	Genes	0.322	0.4203
ZNF548	5	0	0	2	3	0	5	Multiple Exons	0.322	0.4203
OR12D2	4	0	0	1	3	0	4	Genes	0.322	0.3407
SDCBP	4	0	0	1	3	0	4	Genes	0.322	0.3407
ACLY	4	0	0	1	3	0	4	Exon	0.322	0.3407
KCTD10	4	0	0	3	1	0	4	Multiple Exons	0.322	0.3407
SLC28A1	4	0	0	1	3	0	4	Exon	0.322	0.3407
TRAPPC9	4	0	0	1	3	0	4	Genes	0.322	0.3407
FAM151B	4	0	1	0	3	1	3	Multiple Exons	0.322	0.1384
CDC48	4	1	0	0	3	1	3	Exon	0.3018	0.1656
RCC2	4	1	0	0	3	1	3	Exon	0.287	0.1384
MAP2K1	4	1	0	0	3	1	3	Exon	0.287	0.1384
SSR3	4	1	0	0	3	1	3	Genes	0.287	0.1384
CATHL1	7	3	0	0	4	3	4	Genes	0.287	0.0879
CATHL4	7	3	0	0	4	3	4	Genes	0.287	0.0879
EIF2S1	7	2	0	1	4	2	5	Genes	0.2771	0.2312
RNF220	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
ITGAD	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
FKBPL	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
JSP.1	2	0	0	0	2	0	2	Genes	0.2665	0.162
JAM2	2	0	0	0	2	0	2	Exon	0.2665	0.162
COG6	2	0	0	0	2	0	2	Exon	0.2665	0.162
POLR1D	2	0	0	0	2	0	2	Genes	0.2665	0.162
WDR75	2	0	0	0	2	0	2	Exon	0.2665	0.162
TMEM41A	2	0	0	0	2	0	2	Exon	0.2665	0.162
ZHX2	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
ENKUR	2	0	0	0	2	0	2	Exon	0.2665	0.162
AP3B1	2	0	0	0	2	0	2	Exon	0.2665	0.162
BOD1L	2	0	0	0	2	0	2	Exon	0.2665	0.162
C4H7orf62	2	0	0	0	2	0	2	Exon	0.2665	0.162
C8H9orf125	2	0	0	0	2	0	2	Exon	0.2665	0.162
C8H9orf80	2	0	0	0	2	0	2	Exon	0.2665	0.162
CD300A	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
CGRRF1	2	0	0	0	2	0	2	Exon	0.2665	0.162
CHMP2B	2	0	0	0	2	0	2	Exon	0.2665	0.162
CHMP4B	2	0	0	0	2	0	2	Exon	0.2665	0.162
COA5	2	0	0	0	2	0	2	Genes	0.2665	0.162
CYP2D14	2	0	0	0	2	0	2	Genes	0.2665	0.162
DSC1	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
EIF4A2	2	0	0	0	2	0	2	Genes	0.2665	0.162

FABP2	2	0	0	0	2	0	2	Genes	0.2665	0.162
FTMT	2	0	0	0	2	0	2	Genes	0.2665	0.162
GAS7	2	0	0	0	2	0	2	Exon	0.2665	0.162
HDHD3	2	0	0	0	2	0	2	Exon	0.2665	0.162
HIAT1	2	0	0	0	2	0	2	Exon	0.2665	0.162
IP6K3	2	0	0	0	2	0	2	Exon	0.2665	0.162
KCNK2	2	0	0	0	2	0	2	Exon	0.2665	0.162
LIX1L	2	0	0	0	2	0	2	Exon	0.2665	0.162
LYAR	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
MCOLN3	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
MPTX	2	0	0	0	2	0	2	Exon	0.2665	0.162
MRAS	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
NANOG	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
NARG2	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
PDIA3	2	0	0	0	2	0	2	Exon	0.2665	0.162
PGM2L1	2	0	0	0	2	0	2	Exon	0.2665	0.162
PIH1D2	2	0	0	0	2	0	2	Exon	0.2665	0.162
POU2AF1	2	0	0	0	2	0	2	Exon	0.2665	0.162
RALY	2	0	0	0	2	0	2	Genes	0.2665	0.162
RNF38	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
SETMAR	2	0	0	0	2	0	2	Exon	0.2665	0.162
SLC13A4	2	0	0	0	2	0	2	Exon	0.2665	0.162
SNHG12	2	0	0	0	2	0	2	Exon	0.2665	0.162
TERC	2	0	0	0	2	0	2	Exon	0.2665	0.162
TEX14	2	0	0	0	2	0	2	Genes	0.2665	0.162
TMEM66	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
TPD52L2	2	0	0	0	2	0	2	Exon	0.2665	0.162
TTF1	2	0	0	0	2	0	2	Exon	0.2665	0.162
TTF2	2	0	0	0	2	0	2	Multiple Exons	0.2665	0.162
URB2	2	0	0	0	2	0	2	Exon	0.2665	0.162
VSTM1	2	0	0	0	2	0	2	Exon	0.2665	0.162
ZFX	2	0	0	0	2	0	2	Exon	0.2665	0.162
C11H9orf78	2	0	2	0	0	2	0	Exon	0.2665	-0.0056
IFNT	2	0	2	0	0	2	0	Genes	0.2665	-0.0056
IFNT2	2	0	2	0	0	2	0	Genes	0.2665	-0.0056
IFNT3	2	0	2	0	0	2	0	Genes	0.2665	-0.0056
TRPC2	2	0	2	0	0	2	0	Genes	0.2665	-0.0056
UBN1	2	0	2	0	0	2	0	Exon	0.2665	-0.0056
MMRN1	8	2	2	0	4	4	4	Exon	0.2659	0.0448
LHFPL1	8	2	3	0	3	5	3	Genes	0.2512	0.036
GPBP1	4	0	0	2	2	0	4	Exon	0.2241	0.3407
CCDC43	3	0	0	1	2	0	3	Exon	0.1788	0.2548
MAPRE1	3	0	0	1	2	0	3	Multiple Exons	0.1788	0.2548
NDUFV3	3	0	0	1	2	0	3	Exon	0.1788	0.2548
NLRP9	3	0	0	1	2	0	3	Genes	0.1788	0.2548
RNF146B	3	0	0	1	2	0	3	Exon	0.1788	0.2548
RRP36	3	0	0	1	2	0	3	Multiple Exons	0.1788	0.2548
SUB1	3	0	0	1	2	0	3	Exon	0.1788	0.2548

APPENDIX 2.12

FST ANALYSIS OF ENSEMBL CNV GENES

Gene	Total	Angus	Holstein	Brahman	Nellore	Taurus	Indicus	Known/ Novel	Breeds Fst
ENSBTAG00000033169	7	0	3	0	4	3	4	Novel	0.8412
ENSBTAG00000002913	8	0	0	4	4	0	8	Known	0.8236
ENSBTAG00000034378	4	0	0	0	4	0	4	Novel	0.6829
ENSBTAG00000037384	4	0	0	0	4	0	4	Novel	0.6829
ENSBTAG00000026826	4	0	0	0	4	0	4	Novel	0.6829
ENSBTAG00000038245	4	0	0	0	4	0	4	Novel	0.6829
ENSBTAG00000018855	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000039437	4	0	0	0	4	0	4	Known	0.5766
ENSBTAG00000003936	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000003935	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000024240	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000015654	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000000169	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000019020	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000013843	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000017395	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG000000004181	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000026657	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000008137	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000006985	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000007725	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000030529	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000001257	4	0	0	0	4	0	4	Known	0.5766
ENSBTAG00000038720	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000026586	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000025994	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000024633	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000015534	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000018546	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000017560	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000037699	4	0	0	0	4	0	4	Known	0.5766
ENSBTAG00000002726	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG000000005639	4	0	0	0	4	0	4	Known	0.5766
ENSBTAG00000034850	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000038054	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000021059	4	0	0	0	4	0	4	Known	0.5766
ENSBTAG00000019705	4	0	0	0	4	0	4	Known	0.5766
ENSBTAG00000019739	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000033749	4	0	0	0	4	0	4	Novel	0.5766
ENSBTAG00000031517	14	12	0	2	0	12	2	Novel	0.5371
ENSBTAG00000031516	14	12	0	2	0	12	2	Novel	0.5371
ENSBTAG00000031515	14	12	0	2	0	12	2	Novel	0.5371
ENSBTAG00000031160	14	12	0	2	0	12	2	Known	0.5371
ENSBTAG00000001005	14	12	0	2	0	12	2	Known	0.5371
ENSBTAG00000033890	14	12	0	2	0	12	2	Known	0.5371
ENSBTAG00000022339	3	0	0	0	3	0	3	Novel	0.5364
ENSBTAG00000024692	8	0	0	4	4	0	8	Known	0.5294
ENSBTAG00000031100	7	0	0	3	4	0	7	Known	0.5069
ENSBTAG00000031099	7	0	0	3	4	0	7	Known	0.4681
ENSBTAG00000026078	7	0	0	3	4	0	7	Known	0.4681
ENSBTAG00000031097	7	0	0	3	4	0	7	Known	0.4681
ENSBTAG00000031096	7	0	0	3	4	0	7	Known	0.4681
ENSBTAG00000033768	5	0	0	4	1	0	5	Known	0.4673

ENSBTAG00000020385	5	0	0	1	4	0	5	Novel	0.4673
ENSBTAG0000000236	5	0	0	1	4	0	5	Novel	0.4673
ENSBTAG00000040474	5	0	0	1	4	0	5	Novel	0.4673
ENSBTAG00000038995	5	0	0	1	4	0	5	Novel	0.4673
ENSBTAG00000005708	5	0	0	1	4	0	5	Known	0.4673
ENSBTAG00000030630	5	0	0	1	4	0	5	Novel	0.4673
ENSBTAG00000020790	5	0	0	1	4	0	5	Known	0.4673
ENSBTAG00000004082	5	1	0	0	4	1	4	Novel	0.4537
ENSBTAG00000002110	5	1	0	0	4	1	4	Novel	0.4419
ENSBTAG00000016117	5	1	0	0	4	1	4	Known	0.4419
ENSBTAG00000026922	5	1	0	0	4	1	4	Known	0.4419
ENSBTAG00000039322	5	1	0	0	4	1	4	Known	0.4419
ENSBTAG00000002043	5	1	0	0	4	1	4	Known	0.4419
ENSBTAG00000000953	5	1	0	0	4	1	4	Known	0.4419
ENSBTAG00000007184	5	1	0	0	4	1	4	Novel	0.4419
ENSBTAG00000007750	5	1	0	0	4	1	4	Known	0.4419
ENSBTAG00000018446	6	0	0	2	4	0	6	Novel	0.4407
ENSBTAG00000020357	6	0	0	2	4	0	6	Novel	0.4407
ENSBTAG00000023171	6	0	0	2	4	0	6	Known	0.4407
ENSBTAG00000025201	6	0	0	2	4	0	6	Known	0.4407
ENSBTAG00000012026	6	0	0	4	2	0	6	Known	0.4407
ENSBTAG00000013573	6	0	0	2	4	0	6	Novel	0.4407
ENSBTAG00000024691	9	1	0	4	4	1	8	Known	0.4392
ENSBTAG00000026029	9	1	0	4	4	1	8	Known	0.4392
ENSBTAG00000022939	9	1	0	4	4	1	8	Known	0.4392
ENSBTAG00000018854	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000020996	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000018465	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000014461	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000006059	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000033690	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000015794	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000009435	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000014393	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000014791	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000020685	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000015800	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000006253	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000010171	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000013495	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000002081	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000040337	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG000000037613	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000031700	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000021039	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000007644	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000020764	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000019140	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000006831	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000036087	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000005851	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000020959	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000013167	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000027221	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000017492	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000018493	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000038317	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000017352	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG000000023177	3	0	3	0	0	3	0	Known	0.4284
ENSBTAG00000039995	3	0	3	0	0	3	0	Known	0.4284
ENSBTAG00000024545	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000031348	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000008837	3	0	0	0	3	0	3	Novel	0.4284

ENSBTAG0000005208	3	0	3	0	0	3	0	Novel	0.4284
ENSBTAG00000017662	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000020193	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000022715	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000019537	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000034891	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000034921	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000039697	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000034915	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000009197	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000007397	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000003668	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000042421	3	0	0	0	3	0	3	Novel	0.4284
ENSBTAG00000042821	3	0	0	0	3	0	3	Known	0.4284
ENSBTAG00000019588	7	3	0	0	4	3	4	Known	0.4039
ENSBTAG00000009656	7	3	0	0	4	3	4	Known	0.4015
ENSBTAG00000015219	11	1	2	4	4	3	8	Known	0.3953
ENSBTAG00000009000	9	1	0	4	4	1	8	Known	0.3902
ENSBTAG00000030360	8	1	0	3	4	1	7	Known	0.3898
ENSBTAG00000038890	8	1	0	3	4	1	7	Known	0.3898
ENSBTAG00000003182	8	1	0	3	4	1	7	Known	0.3898
ENSBTAG00000000607	6	0	1	1	4	1	5	Novel	0.3797
ENSBTAG00000039520	6	0	0	3	3	0	6	Known	0.3797
ENSBTAG00000038357	6	0	0	3	3	0	6	Known	0.3797
ENSBTAG00000000763	6	0	1	1	4	1	5	Known	0.3797
ENSBTAG00000039763	6	0	1	1	4	1	5	Known	0.3797
ENSBTAG00000040294	6	0	0	3	3	0	6	Known	0.3797
ENSBTAG00000038824	6	0	1	1	4	1	5	Known	0.3797
ENSBTAG00000014238	7	0	2	1	4	2	5	Known	0.3615
ENSBTAG00000009999	11	1	2	4	4	3	8	Known	0.3582
ENSBTAG00000036091	6	1	0	4	1	1	5	Known	0.3565
ENSBTAG00000013626	6	1	1	0	4	2	4	Novel	0.3565
ENSBTAG00000022501	6	1	1	0	4	2	4	Known	0.3565
ENSBTAG00000031166	6	2	0	0	4	2	4	Novel	0.3501
ENSBTAG00000013615	6	2	0	0	4	2	4	Novel	0.3501
ENSBTAG00000009725	7	1	0	2	4	1	6	Known	0.341
ENSBTAG00000030839	3	0	0	0	3	0	3	Known	0.3221
ENSBTAG00000018840	3	0	0	0	3	0	3	Novel	0.3221
ENSBTAG00000011019	3	0	0	3	0	0	3	Known	0.3221
ENSBTAG00000008482	5	0	0	2	3	0	5	Novel	0.322
ENSBTAG00000008862	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000008717	5	0	0	3	2	0	5	Known	0.322
ENSBTAG000000007632	4	0	1	0	3	1	3	Known	0.322
ENSBTAG00000013955	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000034416	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000019910	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000037950	4	0	3	0	1	3	1	Known	0.322
ENSBTAG00000040409	4	0	3	0	1	3	1	Known	0.322
ENSBTAG00000033252	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000002697	4	0	0	3	1	0	4	Novel	0.322
ENSBTAG00000037814	5	0	0	2	3	0	5	Known	0.322
ENSBTAG00000030410	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000036215	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000011262	5	0	0	2	3	0	5	Novel	0.322
ENSBTAG00000013270	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000016740	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000020250	4	0	0	1	3	0	4	Novel	0.322
ENSBTAG00000031832	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000038122	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000031833	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000031835	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000039390	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000040188	4	0	0	1	3	0	4	Known	0.322

ENSBTAG00000039683	4	0	0	1	3	0	4	Known	0.322
ENSBTAG00000032597	8	2	0	2	4	2	6	Novel	0.3088
ENSBTAG00000014326	4	1	0	0	3	1	3	Novel	0.3018
ENSBTAG00000012816	7	1	1	1	4	2	5	Novel	0.3011
ENSBTAG00000039038	7	1	1	4	1	2	5	Known	0.3011
ENSBTAG00000015910	8	2	2	0	4	4	4	Novel	0.2986
ENSBTAG00000018471	4	1	0	0	3	1	3	Novel	0.287
ENSBTAG00000008579	4	1	0	0	3	1	3	Novel	0.287
ENSBTAG00000009444	4	1	0	3	0	1	3	Novel	0.287
ENSBTAG00000033983	4	1	0	0	3	1	3	Novel	0.287
ENSBTAG00000013356	7	3	0	0	4	3	4	Known	0.287
ENSBTAG000000020072	7	3	0	0	4	3	4	Known	0.287
ENSBTAG00000039016	7	1	0	3	3	1	6	Known	0.2831
ENSBTAG00000032787	7	1	1	1	4	2	5	Known	0.2831
ENSBTAG00000038449	7	1	1	1	4	2	5	Known	0.2831
ENSBTAG00000038932	7	2	0	1	4	2	5	Novel	0.2771
ENSBTAG00000040300	7	2	0	1	4	2	5	Novel	0.2771
ENSBTAG00000001945	7	2	0	1	4	2	5	Novel	0.2771
ENSBTAG00000016311	7	2	0	1	4	2	5	Novel	0.2771
ENSBTAG000000023912	7	2	0	1	4	2	5	Known	0.2771
ENSBTAG00000023244	11	1	2	4	4	3	8	Known	0.2724
ENSBTAG00000043157	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000042107	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000043673	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000007444	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000000603	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000007939	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000014724	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000017115	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000001497	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000003002	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000019232	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000012884	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000008367	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000015392	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000015776	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000016982	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000012355	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000032393	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000003864	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000005483	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000024607	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG000000020916	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000037632	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000026501	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000010900	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000017045	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000039431	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000003653	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000005779	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000037583	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000003018	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000020155	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000039069	2	0	2	0	0	2	0	Known	0.2665
ENSBTAG00000034282	2	0	2	0	0	2	0	Known	0.2665
ENSBTAG00000034285	2	0	2	0	0	2	0	Known	0.2665
ENSBTAG00000034289	2	0	2	0	0	2	0	Known	0.2665
ENSBTAG00000005519	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000016750	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000019275	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000000250	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG000000005016	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000012378	2	0	0	0	2	0	2	Known	0.2665

ENSBTAG0000005328	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000017141	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000015805	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000014873	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000020307	2	0	2	0	0	2	0	Novel	0.2665
ENSBTAG00000018710	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000016635	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000008642	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000035333	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000025720	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000015569	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000027412	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000011638	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000021232	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000013387	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000016524	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000000241	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000026309	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000040351	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000037488	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000006282	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000005156	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000024759	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000011549	2	0	2	0	0	2	0	Known	0.2665
ENSBTAG00000000770	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000031030	2	0	0	2	0	0	2	Known	0.2665
ENSBTAG00000012549	2	0	0	2	0	0	2	Known	0.2665
ENSBTAG00000005874	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000008126	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000004407	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000002596	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000003267	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000019107	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000003003	2	0	0	2	0	0	2	Known	0.2665
ENSBTAG00000000812	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000011405	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000006363	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000017196	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000017716	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000004679	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000016456	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000018935	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG000000033603	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000037781	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000020116	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000037619	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000002069	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000019876	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000037750	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000034176	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000001736	2	0	2	0	0	2	0	Known	0.2665
ENSBTAG00000019524	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000013579	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000010241	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000010180	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG000000031410	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG000000007730	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000026992	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000020844	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000024788	2	0	0	2	0	0	2	Known	0.2665
ENSBTAG000000008379	2	0	0	2	0	0	2	Known	0.2665
ENSBTAG00000011122	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000015551	2	0	0	0	2	0	2	Novel	0.2665

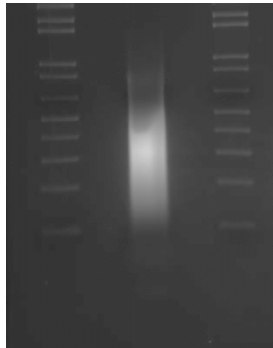
ENSBTAG00000038596	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000037630	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000039665	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000038383	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000038964	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000042475	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000043261	2	0	0	2	0	0	2	Novel	0.2665
ENSBTAG00000042181	2	0	0	2	0	0	2	Novel	0.2665
ENSBTAG00000043315	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000028359	2	0	0	0	2	0	2	Novel	0.2665
ENSBTAG00000042158	2	0	0	0	2	0	2	Known	0.2665
ENSBTAG00000010285	8	2	2	0	4	4	4	Novel	0.2659
ENSBTAG00000032485	8	2	3	0	3	5	3	Novel	0.2512
ENSBTAG00000022175	8	2	3	0	3	5	3	Novel	0.2512

APPENDIX 3.1

SEQUENCING LIBRARY PREPARATION PROTOCOL

➤ Sonication

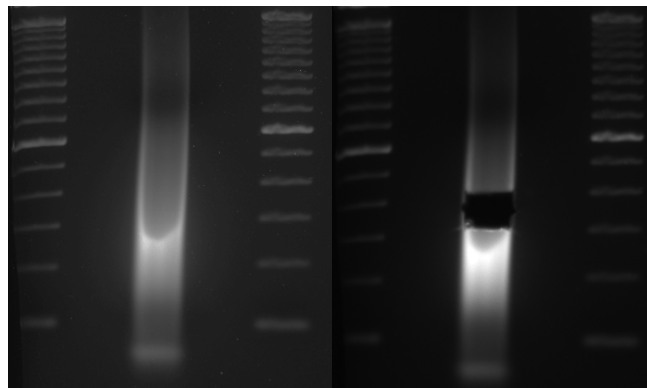
- Sonicate 8 µg of clean genomic DNA diluted in 120 µl elution buffer or H₂O
 - Place tubes on ice during sonication
 - 3-15 second pulses at 12% power, with 30 seconds pauses between pulses
- Test sonication results on gel
 - Run 2% gel at 100V for 1.5 hours with 1Kb+ ladder
 - Visualize gel (Figure 3.1.1)
 - The majority of DNA should range from 200-400 bp



- Perform PCR purification to cleanup sonicated DNA
 - Purelink PCR Kit – Invitrogen
 - Add 400 µl Binding buffer
 - Place on column
 - Centrifuge 10,000xg for 1 minute
 - Pour off liquid
 - Add 600 µl Wash buffer
 - Centrifuge 10,000xg for 1 minute
 - Pour off liquid
 - Add 300µl Wash buffer

- Centrifuge 10,000xg for 1 minute
 - Pour off liquid
 - Spin 2 ½ min at 14,000xg
 - Place column in new 1.7 mL tube
 - Add 35 µl elution buffer
 - Incubate 1min at room temperature
 - Spin 2 ½ min at 14,000xg
 - Nanodrop the collected liquid with the purified DNA
- Proceed to Blunt repair if more than 4 µg DNA in remainder of sample
- Blunt End Repair
 - Reaction components:
 - 45 µl milliQ H₂O
 - 30 µl DNA
 - 10 µl T4 DNA Ligase Buffer w/ 10 mM ATP 10x (NEB# B0202S)
 - 4 µl 10mM dNTPs (Promega U151B)
 - 5 µl T4 DNA Polymerase (NEB# M0203S)
 - 1 µl Klenow Enzyme (large fragment) (NEB# M0210S)
 - 5 µl T4 PNK (NEB# M0201S)
 - Mix reaction gently, spin down
 - Place on thermocycler for 30 minutes at 20°C
 - Purify with Qiagen PCR purification kit – elute in 34 µl EB
- Adenylation
 - Reaction components:
 - 32 µl DNA
 - 5 µl 10X Klenow Buffer (NEB2) (NEB# 7002S)
 - 10 µl 1mM dATPs (NEB# N0440S)
 - 5 µl Klenow 3'5' exo (NEB# M0212S)
 - Mix reaction gently, spin down
 - Place on thermocycler for 30 minutes at 37°C
 - Purify with Qiagen MinElute purification kit – elute in 11.5 µl EB
- Ligation
 - Reaction components:

- 6 μ l H₂O
 - 10 μ l DNA
 - 25 μ l 2X T4 Quick DNA Ligase Buffer (NEB#)
 - 4 μ l Genomic adapter mix (PE adapter) (15 μ M)
 - PE_t_Adapter: (*=Phosphorothioate, HPLC purification of primer)
 - ACACTCTTTCCCTACACGACGCTCTTCCGATc*T
 - PE_b_Adapter: (P- =Phosphate, HPLC purification of primer)
 - P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
 - 5 μ l T4 Quick DNA Ligase (NEB# M0212S)
- Mix reaction gently, spin down
 - Place on thermocycler for 15 minutes at 20°C
 - Purify with Qiagen MinElute purification kit – elute in 18 μ l EB
- Size Selection
- Gel – Pour 2% agarose gel (Certified Low Range Ultra Agarose)
 - 125 mL 1x TAE + 2.5 g Agarose + 1.3 μ l Ethidium bromide
 - Add 4 μ l 5x Loading buffer to ligation product
 - Add 1 kb+ ladder, samples, 1 kb+ Ladder to gel with at least 1 empty lane between each
 - Run at 100 V for 1 ½ hours
 - Visualize gel and record image
 - Size select 3 sizes (250, 350, 425 bp) (gel excision tips, 6.5 mm x 1.0 mm)
 - Gel purify with Qiagen Gel Purification kit
 - Elute in 30 μ l EB



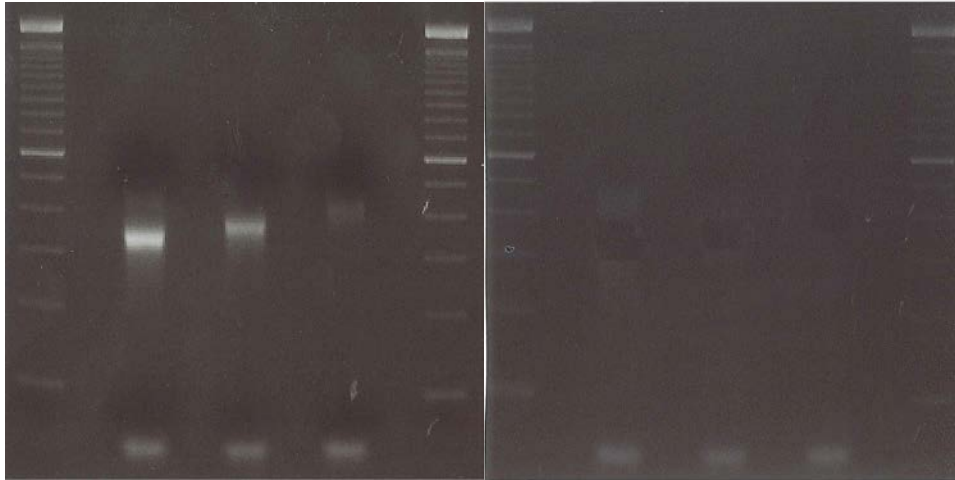
➤ Enrichment

- Reagents:
 - 3 µl ~25 ng DNA
 - 1 µl 25 µM PCR Primer PE1.0: (P- =Phosphate, HPLC purification of primer)
 - AATGATACGGCGACCACCGAGATCTACACTCTTTCCC
TACACGACGCTCTTCCGATC*T
 - 19 µl H₂O
 - 25 µl Phusion DNA Mastermix
 - 1 µl 25 µM PCR Primer PE2.0: (P- =Phosphate, HPLC purification of primer)
 - CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCAT
TCCTGCTGAACCGCTCTTCCGATC*T
- Make 2 reactions per sample
- PCR Settings:
 - 98°C - 30"
 - 98°C - 40" \\\
 - 65°C - 30" --- 12X
 - 72°C - 30" ///
 - 72°C - 5'
 - 4°C - Hold
- Combine replicates and PCR Purify with Minelute – Elute in 20 µl EB

➤ Size selection

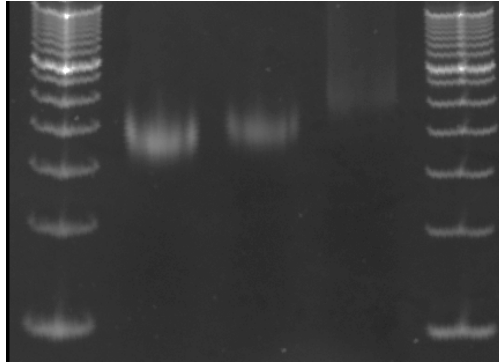
- Gel – Pour 2% agarose gel (Certified Low Range Ultra Agarose)
 - 125 mL 1x TAE + 2.5 g Agarose + 1.3 µl Ethidium bromide
- Add 5 µl 5x Loading buffer to ligation product
- Add 1 kb+ ladder, samples, 1 kb+ Ladder to gel with at least 1 empty lane between each
- Run at 100 V for 1 ½ hours
- Visualize gel and record image
- Size select (see figure below with pre & post selection gel) (gel excision tips, 6.5 mm x 1.0 mm)

- Gel Purify – MinElute – elute in 15 μ l EB



➤ PAGE Confirmation of Size

- 5% Gel:
 - 1.2 mL 10X TBE
 - 2 mL 29:1 Acrylamide/Bis-Acrylamide
 - 8.8 mL H₂O
 - Vortex
 - Add 6.8 μ l Temed
 - 50 μ l 10% APS in H₂O (0.1g + 1mL H₂O)
- Mix 3 μ l library + 5ul H₂O + 3uL 5X Loading Buffer
- Add library and 1 Kb+ ladder to gel and run at 80 V, 45 minutes
- Soak Gel in 80 mL 1X TBE + 9 μ l Ethidium Bromide or 30 min
- Visualize, discard gel
- Estimate library size from gel



- Use the Qubit with HS buffer to determine concentration of library
- Based on length and concentration, dilute to 10 nM in H₂O
- Bioanalyzer – confirm size and concentration

APPENDIX 3.2

SNV CONFIRMATION PRIMERS

Index	Forward Primer	Reverse Primer	Product Length	Gene ID
1	ACTGAATCCAAGACCGTGA	CGTGGTAGGAGGTTTCCAGA	235 bp	<i>GRP75_BOVIN</i>
2	CAGCATTATCAAGCCCAGGT	GGAGGGTGGTAGTGGTGTTT	217 bp	<i>LOC507269</i>
3	CCCACTCCCCTCACACA	CCTCCGGAAATTCTAACGTG	405 bp	<i>ENSBTAG00000032597</i>
4	GAATACCCCATGCTTCAGA	CCACCTGGACACTGGTTAGC	212 bp	<i>SLC35F2</i>
5	CCCTGATTTTGGATGTCTGG	CCCACACCAAATCTGACCAT	289 bp	<i>USP43</i>
6	GAATGACCTGGAATGGGCTA	TTGAGATGCTGTCTCCCTCA	266 bp	<i>USP40</i>
7	AACTCTGGGGGCTACACTGA	ATGGGGGTCTCGAAGGTATC	200 bp	<i>ENSBTAG00000018840</i>
8	CACAACATTGGGACCACAGA	ATGCATGGCCAGATTTTT	235 bp	<i>OR4K17</i>
9	CTCTCAGGCAGGCAGGAC	CTCTTGCTCCCCACATTCC	215 bp	<i>MGC166429</i>
10	GCACAGGAGGCATTGTAGGT	AGCCAGCATAGGAACAGCTC	206 bp	<i>RBM12</i>
11	CATGAGGCATGTGGGATCTA	GAATTGGGATTGCTTTCCAG	241 bp	<i>NEK5</i>
12	CCCCTTCAGTTACTCCACA	CCTGCTTCCTCTCCTCCTTC	205 bp	<i>ITGAD</i>
13	TGCAACAGACCAAGATGAGC	GATCGCAGAGTTGAACACGA	216 bp	<i>DSC2</i>
14	CGCTCTCTCGAGCTCTCTTC	AAGTCCTACAGCCATCCAA	245 bp	<i>GIMAP4</i>
15	CCAGGCACTTGTGTGCAATA	CCCACCCATCTATCCATCTG	253 bp	<i>ENSBTAG00000037840</i>
16	GTCTGACCTCCAGGCTCCTC	CCTCAGAATGGGCCAGATAG	218 bp	<i>O97740_BOVIN</i>
17	CTCCGTGTCTCAGCCTCTTT	CCTCTCCAAGCACCAAAAAC	294 bp	<i>BAT2L</i>
18	GCTTCATCTTGGGAGCTGAG	CCTCTGTGCGAGACCTTCA	205 bp	<i>LOC521950</i>
29	TCTGCTTTTGGTTTGAGCA	GTTGACAAGGCAGCTTCTCC	251 bp	<i>TECRL</i>
20	AGCAGCAATTTTACCCTGT	CAAGAAACGTGCTGCCTATG	206 bp	<i>XIRP2</i>
21	CCAGCCAACAATGGACTT	TTCAGGTGTCGTTCAAGGAA	200 bp	<i>HIAT1</i>
22	CCCAAAGGAAGAAGTCGATG	CCATGTCCTTTTTCCCCTCT	258 bp	<i>RCN2</i>
23	GGATCGCACAAACCAATAACC	GCACACATCTGGCTGTTCTG	213 bp	<i>ENSBTAG00000011932</i>
24	TGCTCAGGATGGTGGTGATA	CTTTCTGGACGCACTCATCA	227 bp	<i>OR8S1</i>
25	GTGGCTGGGGAGGAAGTAAT	CCTTTGCCACATCTGGAGTT	214 bp	<i>ENSBTAG00000032145</i>
26	TGGCTTTATCCTAATCGTAGCC	GCAGCCGATAAAGAAAATCA	260 bp	<i>UBE2D3P</i>
27	CCCTGATGTACCCCTTCT	GTGTCGCTCTCGTGCAGTAA	235 bp	<i>RCN3_BOVIN</i>
28	CCCTGAAGTTCCTCCCAACT	CCGCTCAAGTTTTTCAGAAG	286 bp	<i>BTN3A2</i>
29	CTGCCAAGAACCATGTGATG	TGACAAGGTTCCGTTATCCTG	261 bp	<i>FZD3</i>
30	CATCCCTGATTGTCCTTTTCA	CGTACACGTCCCCATAAGAAA	245 bp	<i>ENSBTAG00000021830</i>

APPENDIX 3.3

BIOLOGICAL PROCESS CLASSIFICATIONS

Please see attached Microsoft Excel file for Appendix 3.3 containing a list of all biological process terms and manually curated functional groups.

APPENDIX 3.4

ANGUS AND NELLORE SNV ANALYSIS WITH MINIMUM DEPTHS OF 5 TO 10X

Please see attached Microsoft Excel file for Appendix 3.4 complete SNV annotations against ensembl genes. The analyses are per sample with minimum read-depths ranging from 5 to 10. SNVs are annotated based on their genic locations and on predicted effects on amino acids. The annotations are divided into 24 tables, 2 for each sample at each read-depth.

APPENDIX 3.5

GENES WITH DN/DS > 1 IN ANGUS AND NELLORE

Genes dN/dS >1			
Angus		Nellore	
<i>PRG4</i>	3.2347	<i>AOX1</i>	2.0988
<i>GIMAP4</i>	2.6353	<i>CD163L1</i>	1.7441
<i>PARK2</i>	1.6939	<i>NEIL3</i>	1.6624
<i>GIMAP7</i>	1.4697	<i>STOX1</i>	1.6376
<i>MCM2</i>	1.408	<i>DSG2</i>	1.583
<i>SLC26A2</i>	1.3159	<i>HAVCR2</i>	1.4066
<i>GSDMB</i>	1.2306	<i>LOC786254</i>	1.38
<i>MRPL2</i>	1.2234	<i>LRRC6</i>	1.3682
<i>KIR2DS1</i>	1.1544	<i>LOC534155</i>	1.3358
<i>UBE2D3</i>	1.1081	<i>LOC100125266</i>	1.2408
<i>GIMAP5</i>	1.0165	<i>RPS3</i>	1.2372
		<i>C16H1orf170</i>	1.2206
		<i>RTTN</i>	1.1945
		<i>DPYD</i>	1.1859
		<i>COX20</i>	1.1798
		<i>ZBTB40</i>	1.1191
		<i>MYT1</i>	1.1161
		<i>ZSCAN26</i>	1.1037
		<i>ZNF280A</i>	1.0851
		<i>UBE2B</i>	1.074
		<i>ITGB1BP2</i>	1.0684
		<i>NBAS</i>	1.0543
		<i>ADGB</i>	1.0187
		<i>SGOL2</i>	1.0165
		<i>TTF2</i>	1.0134
		<i>ALPK2</i>	1.0119
		<i>ULBP3</i>	1.0038

APPENDIX 4.1

CHIP ANNOTATION

Please see attached Microsoft Excel file for Appendix 4.1 containing complete annotations of H₃K₄me₃ regions. The analyses include annotations using both RefSeq and ensembl gene IDs. Summary tables are included for regions overlapping CpG islands, conserved regions, tandem repeats, CNVs, INDELS, SNVs, and regions enriched with DNA methylation. The annotations are divided into 8 tables: H₃K₄me₃, H₃K₄me₃ overlapping CpG islands, H₃K₄me₃ overlapping Conserved regions, H₃K₄me₃ overlapping CpG islands and conserved regions, H₃K₄me₃ overlapping DNA methylation, H₃K₄me₃ overlapping CpG islands and DNA methylation, H₃K₄me₃ overlapping Conserved regions and DNA methylation, H₃K₄me₃ overlapping CpG islands, conserved regions and DNA methylation.

APPENDIX 4.2

DNA METHYLATION ANNOTATION

Please see attached Microsoft Excel file for Appendix 4.2 containing complete annotations of DNA methylation regions. The analyses include annotations using both RefSeq and ensembl gene IDs. Summary tables are included for regions overlapping CpG islands, conserved regions, tandem repeats, CNVs, INDELS, SNVs, and regions enriched with DNA methylation. The annotations are divided into 2 tables: DNA methylation regions, DNA methylation regions not overlapping H₃K₄me₃.

APPENDIX 4.3

CHIP AND METHYLATION SNV DENSITIES AND ORS

Please see attached Microsoft Excel file for Appendix 4.3 containing complete annotations SNV densities and ORs overlapping genomic regions including intergenic, intronic, exonic, 3' UTR, and PPRs. SNV ORs were calculated for all regions in the presence of: H₃K₄me₃; H₃K₄me₃ overlapping CpG islands; H₃K₄me₃ overlapping conserved regions; H₃K₄me₃ at non-conserved regions; H₃K₄me₃ overlapping CpG islands and conserved regions; H₃K₄me₃ overlapping DNA methylation; H₃K₄me₃ overlapping CpG islands and DNA methylation; H₃K₄me₃ overlapping conserved regions and DNA methylation; H₃K₄me₃ overlapping CpG islands, conserved regions and DNA methylation; DNA Methylation overlapping CpG islands; DNA Methylation overlapping conserved regions; and DNA Methylation H₃K₄me₃ overlapping CpG islands and conserved regions

APPENDIX 5.1

PLINK ANALYSIS BATCH FILE

```
cd C:\plink-1.07-dos
```

```
plink --file BovineHD\BovineHD_9_17_2011 --make-bed --cow
```

```
plink --bfile BovineHD\BovineHD_9_17_2011 --blocks --cow
```

```
plink --bfile BovineHD\BovineHD_9_17_2011 --hap plink.blocks --hap-freq --cow
```

```
plink --bfile BovineHD\BovineHD_9_17_2011 --het --cow
```

```
plink --bfile BovineHD\BovineHD_9_17_2011 --homozyg --cow
```

```
plink --file BovineHD/BovineHD_9_17_2011 --homozyg --homozyg-group --cow
```

APPENDIX 5.2

RUNS OF HOMOZYGOSITY

Please see attached Microsoft Excel file for Appendix 5.2 containing 692 runs of homozygosity identified by the BovineHD SNP Beadchip in four Angus and four Nellore cows. All locations are in respect to the bovine Umd3 genome assembly.

APPENDIX 5.3

BP ANALYSIS OF IMPUTED AND NON-IMPUTED CNV GENES

BP of Imputed CNV Genes			
Term	Genes	Fishers Method	p value
Immunity and Defense	34	25.06828014	4.87442E-05
Signal Transduction	79	33.16667437	6.27983E-08
Sensory Perception	79	43.11128085	2.33069E-09
Regulation of Cell Cycle	-	-	-
Cellular Process	4	2.548042538	0.110430998
Metabolism	5	3.288467146	0.069768127
Protein processing	3	3.209985472	0.073190133
Nucleic Acid Metabolism	-	-	-
Miscellaneous	6	4.812474773	0.090153871
Developmental Process	-	-	-
BP of Non-Imputed CNV Genes			
Term	Genes	Fishers Method	p value
Immunity and Defense	8	7.22996575	0.026917386
Signal Transduction	-	-	-
Sensory Perception	-	-	-
Regulation of Cell Cycle	-	-	-
Cellular Process	-	-	-
Metabolism	-	-	-
Protein processing	-	-	-
Nucleic Acid Metabolism	-	-	-
Miscellaneous	-	-	-
Developmental Process	-	-	-