IDENTIFICATION OF COPY NUMBER VARIANTS IN THE NELLORE AND ANGUS FOUNDERS OF A BEEF CATTLE MAPPING POPULATION AND THEIR EFFECTS ON GROWTH AND PRODUCTION TRAITS

A Dissertation

by

YUE XING

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Clare Gill
Committee Members,	James Cai
	Andy Herring
	David Riley
	Jason Sawyer
Head of Interdisciplinary Program,	David Threadgill

December 2018

Major Subject: Genetics

ABSTRACT

Copy number variants (CNV) are insertions or deletions of 1 kb or larger in a genome with variable number of copies compared to a reference genome that can affect phenotypic expression. Methods for identifying and applying CNV are less well developed than those for single nucleotide polymorphisms (SNP). Because CNV can encompass genes or their regulatory regions and contribute to genetic variation of traits of economic importance in beef cattle, it is of interest to study their effects; for example, on birth and weaning weights. This study identified and characterized bovine CNV in founders of a beef cattle mapping population, compared the performance of CNV identification methods, proposed ways to obtain CNV sets with fewer false discoveries, developed an approach to use SNP having high linkage disequilibrium with CNV to analyze association of CNV to economically important traits using genome-wide association studies (GWAS), and developed approaches to incorporate CNV into genomic selection for economically important traits.

The performance of read-pair based methods highly rely on the depth of coverage of the tested genome compared to the control genome, selection of a control animal, and selection of window size. Using the consensus set of CNV regions (CNVR) from different control animals may lower the false discovery rate. Read-pair and split read based methods were relatively more stable, but could not identify large insertions. Split read based methods also had difficulty identifying other kinds of large-scale structural variants. Because any method alone was not comprehensive enough, and may

result in a high false discovery, it was better to focus on combined methods and the common set of CNVR. GWAS identified the association of CNVR with birth and weaning weights, and predictive modeling helped phenotype prediction by CNVR. Random forest and Bayesian sparse linear mixed models were the best models with highest prediction accuracy. The additive SNP model had slight advantages over dominance and recessive SNP models. Some novel genes that may have effects on birth and weaning weight were discovered. Further analysis will be required to determine if the gene effects discovered are real and how they affect these traits.

DEDICATION

To my family

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Gill, and my committee members, Dr. Cai, Dr. Herring, Dr. Riley, and Dr. Sawyer, for their guidance and support throughout the course of this research.

Thanks to my committee chair, Dr. Dabney, and my committee members, Dr. Gill, and Dr. Longnecker, of my secondary degree, Master of Science in Statistics, for their guidance and support for Chapter V of this research.

Thanks to the Texas A&M University high performance research computing system (HPRC), and Texas A&M Institute for Genome Sciences and Society High Performance Compute Cluster (TIGSS HPC Cluster), for computational resources and systems administration to conduct this research.

Thanks to my lab mates for their support to conduct this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my mother and father for their encouragement and to my husband for his patience and love.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Clare Gill, James Cai, Andy Herring, David Riley, and Jason Sawyer of the Interdisciplinary Program of Genetics.

The sequence alignment data used in Chapter II was provided by graduate students enrolled in Applied Animal Genomics taught by Clare Gill. The discussion about CNV-seq in Chapter II was helped by discussion with J.-W. Choi. The validation analyses depicted in Chapter II and III were conducted in part by Xiao Li of the Department of Molecular and Cellular Medicine.

The analyses depicted in Chapter V was supervised by Professors Alan Dabney and Michael Longnecker of the Department of Statistics, and Clare Gill of the Department of Animal Science.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by Charles Robertson Fellowship of the Department of Animal Science, teaching assistantships from the Department of Biochemistry and Biophysics and the Department of Biology, and Texas A&M AgriLife Research through an Enhancing Research Capacity for Beef Production Systems grant awarded to Clare Gill.

NOMENCLATURE

AS	Assembly
ASIP	Agouti signaling protein
BIC	Bayesian Information Criterion
BSLMM	Bayesian sparse linear mixed model
BVD1.18	T cell receptor delta chain variable region BVd1.15
CHORDC1	Cysteine and histidine-rich domain-containing protein 1
CNV	Copy number variants
CNVR	Copy number variant regions
CV	Cross-validation
GBLUP	Genomic best linear unbiased prediction
GO	Gene ontology
GWAS	Genome-wide association studies
IBD	Identical-by-decent
LD	Linkage disequilibrium
LMM	Linear mixed model
MLR	Multivariate linear regression
MSE	Mean squared error
PTPRT	Protein tyrosine phosphatase, receptor type T
PVE	Proportion of variance in phenotype explained by a given SNP
qPCR	Quantitative polymerase chain reaction

QTL	Quantitative trait loci
RCN	Relative copy number
RD	Read depth
RF	Random forest
RNF122	Ring finger protein 122
RP	Read-pair
RT	Regression tree
SNP	Single nucleotide polymorphisms
SR	Split read
SV	Structural variants
WGS	Whole genome sequences

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

CONTRIBUTORS AND FUNDING SOURCES vi

NOMENCLATURE vii

TABLE OF CONTENTS ix

LIST OF FIGURES xi

LIST OF TABLES xiii

CHAPTER I INTRODUCTION 1

I.1 Literature Review 2
I.2 Approach 6

CHAPTER II IDENTIFICATION OF COPY NUMBER VARIANTS IN THE NELLORE AND ANGUS FOUNDERS OF A BEEF CATTLE MAPPING POPULATION 12

II.1 Introduction 12
II.2 Methods 14
II.3 Results and discussion 18
II.4 Conclusions 34

CHAPTER III COMPARISON OF THREE SOFTWARE APPLICATIONS FOR COPY NUMBER VARIANT DETECTION AND THE ALGORITHMS BEHIND THEM 36

III.1 Introduction 36
III.2 Methods 41
III.3 Results and discussion 43
III.4 Conclusions 58

CHAPTER IV GENOME-WIDE ASSOCIATION STUDY OF SNP-TAGGED COPY NUMBER VARIANT REGIONS IN A BEEF CATTLE MAPPING POPULATION 60

IV.1 Introduction 60
IV.2 Materials and Methods 62
IV.3 Results and discussion 64
IV.4 Conclusions 73

CHAPTER V PREDICTION OF EFFECTS OF CNVR ON BIRTH AND WEANING WEIGHT IN A BEEF CATTLE MAPPING POPULATION 78

V.1 Introduction 78
V.2 Methods 81
V.3 Results and discussion 87
V.4 Conclusions 108

CHAPTER VI CONCLUSION 110

REFERENCES 113

APPENDIX A FIGURES 140

APPENDIX B TABLES 155

APPENDIX C ADDITIONAL FILES 165

LIST OF FIGURES

Page

Figure II.1 Comparison of CNVR counts and nucleotide content by window size…….. 21

Figure II.2 Effect of window size on detection of CNVR in an Angus and a Nellore……...……22

Figure II.3 $\log_2$ ratios for CNVR across the genome for different window sizes…….23

Figure II.4 CNVR identified in Nellore and Angus with a minimum length of 25 kb…….24

Figure II.5 Impact of different control animals on identification of CNVR…….26

Figure II.6 CNVR detected in common among Angus and Nellore……..27

Figure II.7 CNVR identified in Nellore and Angus by CNV-seq and in common with previous studies…….28

Figure II.8 Number of QTL overlapping a consensus set of CNVR…….32

Figure II.9 $\log_2$ ratios for putative CNVR 4_27 on BTA 4 from 75026251-75116250 bp……..35

Figure III.1 The major algorithms used in CNV identification for NGS data……37

Figure III.2 Type and number of CNVR identified by BreakDancer, CNV-seq and RAPTR-SV……...46

Figure III.3 Comparison of number and size for CNVR identified by BreakDancer, CNV-seq and RAPTR-SV……48

Figure III.4 CNVR distribution by BreakDancer, CNV-seq and RAPTR-SV across the genome……...50

Figure III.5 CNVR identified by BreakDancer, CNV-seq and RAPTR-SV between 60 mb to 70 mb on bta13 in Nellore N01…….52

Figure III.6 Venn Diagram of CNVR shared between different software applications …….54

Page

Figure IV.1 Circos plots of $-\log_{10} P-$ values for CNV tagging SNP……67

Figure IV.2 Manhattan plots of $-\log_{10} P-$ values for SNP……69

Figure V.1 Flow chart of the study……84

Figure V.2 MSE of each CNVR from CV with the four models……..88

Figure V.3 MSEs of varying numbers of ranked CNVR from CV……...89

Figure V.4 Best CNVR to obtain minimum MSE for MLR model……...92

Figure V.5 Prediction in testing data set using MLR model……99

Figure V.6 The common genes identified by the four models……..……..104

LIST OF TABLES

Page

Table II.1 Proportion of CNVR mapped to RefSeq genes……..……29

Table II.2 GO terms that are significantly enriched in both Nellore and Angus animals……..31

Table II.3 Correlations of log2 ratios calculated by CNV-seq to RCNs for tested CNVR……..33

Table III.1 Types of structural variants identified by BreakDancer, CNV-seq and RAPTR-SV……...45

Table III.2 Comparison of length of CNVR identified by BreakDancer, CNV-seq and RAPTR-SV…….……..47

Table III.3 Distribution of CNVR shared by all three software applications in nellore and angus breeds…….55

Table IV.1 The SNP sets used in this study……...65

Table IV.2 Number of RefSeq genes overlapping with significant SNP in the 6 CNVR-SNP sets…….72

Table IV.3 The common genes discovered overlapping with CNVR- tagging SNP in two of the CNVR-SNP sets…….74

Table IV.4 Comparison of SNP on BTA29 close to CHORDC1 identified by Anton et al. and our study……..75

Table IV.5 RefSeq genes overlapping with significant SNP that have direct overlap with CNVR……...76

Table V.1 Number of ranked CNVR to obtain minimum MSE for BSLMM model……...90

Table V.2 Number of ranked CNVR to obtain minimum MSE for RT model……...93

Table V.3 Number of ranked CNVR to obtain minimum MSE for RF model……...95

Table V.4 Prediction in testing data set using BSLMM, RT and RF models……..97

Table V.5 Number of RefSeq genes overlapping with best collections of CNVR for each model……102

CHAPTER I

INTRODUCTION

Copy number variants (CNV) are insertions or deletions of 1 kb or larger in a genome that are present in a variable number of copies in comparison to a reference genome (reviewed by [1]). The reference genome may be the genomic reference for a species (e.g. UMD3.1 bovine reference sequence) or it may be a reference (control) genome for the experiment. Although CNV account for more nucleotides of genetic variation in a mammalian genome than single nucleotide polymorphisms (SNP) [2, 3], the methods for identifying and applying CNV are less well developed than those for SNP. Because CNV can encompass genes or their regulatory regions [3-6], I expect CNV will contribute to genetic variation of traits of economic importance in beef cattle. The goal of this study is to identify bovine CNV in a *Bos taurus indicus* x *Bos taurus taurus* (Nellore-Angus) cross population and to develop an approach to incorporate CNV into genome-wide association studies (GWAS) and genomic selection for growth and production traits.

The following specific objectives are proposed to achieve the goals of this study: 1) To compare the performance of software packages designed for CNV discovery; 2) To identify and characterize CNV from Angus and Nellore cattle; 3) To determine the proportion of bovine CNV captured with SNP data; 4) To adapt genome-wide association analysis methods for use with CNV and then identify CNV associated with

growth and production traits in a Nellore-Angus mapping population; 5) To develop a method to incorporate CNV into genomic selection.

I.1 Literature Review

Copy number variation can arise both meiotically and somatically [7] and is the primary mode by which individuals accumulate mutations. The mutation rate for CNV is 1.7 x 10-5 per locus [4] compared with 1.8 x 10-8 for SNP [8]. The median size of CNV in humans is 2.9 kb [9] and it is estimated that a typical genome has ~160 copy number variants [3] encompassing 5 to 24 Mb of the genome [10]. The distribution of CNV is nonrandom and strongly correlated with exons, segmental duplications, and transposable elements [11]. Mechanisms known to cause CNV are non-allelic homologous recombination, non-homologous end-joining, the break-fusion-bridge cycle, and replication errors following template-switching or fork stalling (reviewed by [12]). In humans and mice, CNV have been shown to account for 18 to 74% of the genetic variation in gene expression, dependent on the tissue, and at least part of the differential expression is attributable to gene dosage effects [13, 14]. Large multiallelic CNV, which are a small subset of all CNV, cause 88% of the variation in human gene dosage [15].

Because CNV can impact phenotypic expression [3], there has been considerable interest in systematically identifying CNV genome-wide in cattle to incorporate variation attributable to CNV into selection programs. Matukumalli et al. [16] used the Illumina BovineSNP50 assay to detect 79 homozygous deletions in samples from diverse *Bos taurus*, *Bos indicus* and composite breeds previously characterized by the Bovine Hapmap Consortium [17]. Bae et al. [18] also used the 50K chip and found 368

CNV regions (CNVR) with median 171 kb size, and in a survey of 2654 bulls from Italian beef and dairy breeds, Cicconardi et al. [19] identified 326 CNVR with about 60% of the CNV overlapping those identified by [18]. Similar to humans, the distribution of CNV in cattle is nonrandom with 20 to 25% of bovine CNV in segmental duplications [20, 21]. However, Hou et al. [21] showed that some SNP in CNVR have been depleted from the commercial assays, so SNP-based detection approaches are expected to underestimate the true number of CNV in the cattle genome. Recently, Xu et al. [22] used the BovineSNP50 assay to identify CNV associated with milk production traits in Holsteins. They performed a conventional genome-wide association study with SNP then characterized linkage disequilibrium between SNP and CNV to infer 34 CNV significantly associated with at least one milk trait.

Array comparative genomic hybridization (CGH) is another approach to detect CNV. Liu et al. [23] used Roche NimbleGen 385,000 probe whole-genome CGH arrays to discover 177 high-confidence CNVR in 17 breeds of cattle. Using the same assay, Kijas et al. [24] identified CNVR spanning 0.45% of the genome with a minimum detectable length of 80 kb in a small sample of Angus, Brahman and composite cattle. To increase the resolution of CNV detection, Fadista et al. [20] designed a 6.3 million probe NimbleGen CGH array. Identified CNVR ranged from 1.7 kb to 2 Mb (median 16.7 kb), and spanned 23 Mb of the bovine genome. However, in 2012 Roche discontinued manufacturing arrays and these assays are no longer commercially available.

In most of these earlier cattle studies, only large CNVR were found, probably reflecting the low resolution of these probe-based methods. Discovery of CNV in bovine whole-genome sequences (WGS) from Angus, Chikso, Hanwoo, Holstein, Jeju Heugu, and Nellore bulls, has been described recently [25-29]. The WGS in these studies ranged from low coverage (4x) for the Holstein [26] and Jeju Heugu [28], to high coverage (13-22x). In most of these studies, either CNV-seq [30] or mrFAST and mrsFAST [31] were used to detect CNV based on read depth. The hypothesis for read depth methods is that the depth of coverage of a genomic region is correlated with copy number of the region [32]. However, these tools are designed for pairwise case/control comparisons rather than for characterizing populations and can have very high false discovery rates. Most recently, Shin et al. [29] identified deleted CNV in cattle using Genome Strip [33], which was developed to identify CNV for the 1000 human genomes project [3, 15, 34, 35]. In addition to read depth, Genome Strip uses read-pair information to ascertain deletion alleles from read-pairs that map further apart than expected based on the insert size distribution of the library of DNA fragments for sequencing. Genome Strip also considers that a true polymorphism creates heterogeneity in a population, but this feature is sensitive to depth of coverage of the sequenced individuals.

Stothard et al. [25] found 790 CNV ranging from 1,841 bp to 28,029 (median 3,171 bp) and covering 3.3 Mb of autosomal sequence. CNV were all described as gains with respect to breed (i.e. gains in Angus vs. gains in Holstein). Bickhart et al. [26] identified 1265 CNV covering 55.6 Mb of sequence. Only CNV calls > 10 kb in length were declared (average 49.1 kb) and by validation with qPCR and array CGH the false

discovery rate was estimated as 8.1%. There were fewer shared CNV in pairwise comparisons of the Nellore bull and the Angus or Holstein bulls than among the taurine bulls, suggesting that there is greater CNV diversity between the subspecies than across taurine breeds. Choi et al. [27] used more conservative criteria for cnv-seq than [25] to reanalyze the Angus and Holstein sequences from [25] and compared those to Hanwoo. The CNVR ranged from 5,770-35,104 bp (median 7,178 bp) for the Angus vs. Hanwoo comparison and 4,176-22,398 bp (median 7,472 bp) for the Holstein vs. Hanwoo comparison. Choi et al. [28] used the same criteria for CNV-seq to identify CNV in other Korean breeds (Chikso, Jengu Huegu). In pairwise comparisons against Korean Holstein, they found 992 CNVR (median length 13,780 bp) covering 16 Mb of the genome for Hanwoo, 1881 CNVR (median 9,156 bp) spanning 4.7 Mb for Jeugu Heugu, and 1,881 CNVR (median 13,626bp) over 30.8 Mb for Chikso. There were appreciably more CNVR gains in Holstein than in Hanwoo and Chikso, but not Jengu Heugu. This may be attributed to introgression of European breeds of cattle into Jengu Heugu [28]. There was a tendency for CNVR to be nonrandomly distributed across the chromosomes with more CNVR near the telomeres. Shin et al. [29] simultaneously characterized 10 Holsteins and 22 Hanwoo using Genome Strip, which had sufficient power to detect deleted CNV but not insertion events. A total of 6,811 deleted CNV (20% FDR) with an average length of 2.7 kb covering 18.6 Mb of autosomal sequence were detected.

In addition to genome-wide efforts to identify CNV in cattle, some researchers have focused on characterizing specific regions of the genome known to be associated with traits of economic importance. For example, Xu et al. [36] found that CNV

including the molecular interacting CasL-like protein 2 (MICAL-L2) gene are associated with body height, weight and length of cattle. Durkin et al. [37] showed that color sidedness is caused by a complex CNV of homologous but non-syntenic allele. One allele results from the translocation to BTA 29 of a 492-kb region of BTA 6 encompassing viral oncogene homolog (KIT). The second allele is a 575 kb region fused with v-kit Hardy-Zuckerman 4 feline sacoma KIT on BTA 6 that is derived from the BTA29 allele. Subsequently, Venhoranta et al. [38] showed that the CNV allele on BTA29 causes gonadal hypoplasia in white cattle. Zhang et al. [39] identified two CNVR associated with body measurements in Chinese cattle, and demonstrated one of the CNVR had significant negative effects on expression of the PLA2G2D gene. McDaneld et al. [40] showed that a region of chromosome 5 associated with decreased reproductive efficiency in Bos indicus-influenced females contains a deletion CNV. These studies suggest that there is scope to consider CNV for genomic selection in cattle breeding to improve the growth and beef production related traits in cattle breeds.

I.2 Approach

I.2.1 Sequence Data

Aligned sequence data from Gill et al. (unpublished) will be used for CNV discovery. Briefly, these bam-formatted alignment files are based on Illumina paired-end 100 bp sequence from 7 Nellore (*Bos taurus indicus*) bulls and 6 Angus (*Bos taurus taurus*) cows that were the founders of a multigenerational mapping population [41]. After quality control trimming with fastq-mcf [42], reads were aligned to the UMD3.1 bovine assembly [43] with BWA [44], and local realignment and recalibration of quality

scores was done using GATK 3.2 [45]. Genome coverage ranges from 33x to 88x per animal.

I.2.2 Comparison of Software for Detection of Copy Number Variants

The performance of several packages for the detection of CNV in massively parallel sequence data will be compared. CNVer [46] combines read depth and read pair information to detect CNV gains and losses. It is distributed with a package for the analysis of human data, so a bovine package will be built for this tool. BreakDancer [47], is designed to detect structural variants including insertions, deletions, inversions and translocations from paired-end read mapping [48]. Several studies have used it to detect gene rearrangements, chromosomal translocations and inversions in the genome [49, 50]. However, BreakDancer applies a hard cluster method requiring reads map to a single genomic location, so it cannot detect variants in repetitive regions [51]. For the current study, only insertions and deletions will be considered. RAPTR-SV [52] combines read pair and split read data to detect insertions, deletions and tandem repeats in the genome. Because other bovine researchers have used CNV-seq [30], I will benchmark performance of the other packages against CNV-seq. The read depth method is better able to ascertain large insertions and deletions in complex regions [53]. Computational speed and concordance of CNV among call sets will be considered. For those packages that require a control sequence for comparison, I will use one of the Angus sequences and copy numbers will be reported relative to the control [51].

CNV-seq [30], in specific, can utilize the read coverage of the sequencing data and calculate the best window size which makes copy ratios between case and control

significantly differ. Then CNV-seq models the number of short reads in a genomic region by a Poisson distribution which, however, might not be an optimal model in some situations [54].

I.2.3 Optimization of CNV Detection with CNV-seq

CNV-seq [30], which identifies CNV based on differences in read depth after normalization for depth of coverage across the genome, will be used to identify CNVR. The Angus cow that has the highest coverage and is most unrelated to the other animals will be chosen as the control individual for all pairwise comparisons. To enable comparison to previous studies [25, 27, 28] detecting bovine CNV using cnv-seq, strict threshold values (P = 0.001 and log2 threshold = 0.7) will be applied. Five combinations of overlapping windows and consecutive windows for annotation will be used for comparison and annotation: 1 kb overlapping window with 4 consecutive windows, 5 kb overlapping window with 4 consecutive windows, 3 kb overlapping window with 10 consecutive windows, 4 kb overlapping window with 10 consecutive windows, and 5 kb overlapping window with 10 consecutive windows. For these approaches the minimum detectable CNV will be 2 kb, 10 kb, 15 kb, 20 kb, and 25 kb, respectively. Identified CNV will be summarized across the genome and by chromosome. To evaluate the influence of choice of control animal on CNV detection, analysis will be repeated with a Nellore as the control sequence. The effect of depth of coverage of the control animal will also be tested using a different Angus and a different Nellore as the control.

I.2.4 Summarization and Visualization of CNVR

To summarize the CNVR data, the genome will be split in to consecutive windows of 50 kb and CNVR from each animal will be assigned to a window by their length and position. If a CNVR is longer than 50 kb, it will be split by the window. The total number of CNVR from Nellore and Angus falling in each window will be calculated. CNVR will be plotted as log2 values by position on the chromosome. Positive $\log_2$ values represent insertions (gains) with respect to the control, whereas negative $\log_2$ values represent deletions (losses). All graphs will be plotted using R. Graphs summarizing CNVR for the whole genome will be generated from a modified script for Manhattan plots.

I.2.5 Annotation of CNVR

Perl scripts will be written to map the CNV to the bovine RefSeq genes downloaded from the UCSC website. If a CNVR does not hit a gene, the genes in the neighboring upstream or downstream 50 kb window will be identified because the CNVR may function in the regulatory region of a gene.

I.2.6 Gene Ontology Enrichment Analysis

DAVID [55] will be used for GO enrichment analysis. Lists of genes overlapping with CNV in each of the animals will be analyzed. Within category Benjamini-Hochberg correction [56] will be applied to control the false discovery rate.

I.2.7 qPCR Validation of CNVR

Quantitative PCR will be used for validation. Twenty CNV be randomly chosen from the set of CNV overlapping with genes. As in prior studies [21, 27], we will use BTF3 for normalization, because it has not been found in a CVNR in cattle and is

assumed to have only two copies. Primers will be designed using Primer 3 Plus and applying the design criteria of [21, 57]. In particular, the GC clamp will be set to 2, self-annealing will be minimized and the Tm for a pair of primers will be ~60°C. Standard qPCR procedures for SYBR green detection with technical triplicates will be used. The same Angus cow used as the control animal for CNV-seq will be the calibrator for qPCR. Relative copy number for each sample will be calculated as $2^{-\Delta\Delta C_t}$.

I.2.8 Linkage Disequilibrium between CNV and SNP

SNP within 1 Mb of detected CNV will be extracted from imputed genotypes (Gill et al., unpublished) for the McGregor Genomics Cycle 1 population [41]. Using a SNP centered on the position of the CNV, linkage disequilibrium (r^2) between each SNP against the CNV will be calculated using PLINK [58] to assess how effectively SNP data tag CNV.

I.2.9 Genome-wide Association Study with CNV

If I demonstrate that SNP can serve as an effective proxy for most CNV, then I will perform a SNP-based GWAS for growth and production traits using the linear mixed model approach implemented in GEMMA [59] that incorporates the genomic relationship matrix to account for relatedness and population stratification. The coordinates of SNP that are significant after multiple testing correction will be compared to the coordinates of detected CNV to infer that the CNV is associated with the trait as in [22]. However, if SNP cannot serve as a proxy or there is a subset of CNV not captured by SNP, then the statistical methods developed by Barnes et al. will be adapted accomplish the CNV GWA studies.

I.2.10 Genomic Selection with CNV

If SNP are an adequate proxy for CNV then standard Bayesian procedures such as those implemented in GenSel will be applied [60]. Otherwise machine learning processes will be used for genetic prediction. It is likely that this will require that a scoring system is developed to assign copy number for CNV and that CNV haplotypes are established following the approaches described in [3].

CHAPTER II

IDENTIFICATION OF COPY NUMBER VARIANTS IN THE NELLORE AND ANGUS FOUNDERS OF A BEEF CATTLE MAPPING POPULATION

II.1 Introduction

CNV are insertions or deletions of 1 kb or larger in a genome that are present in a variable number of copies in comparison to a reference genome (reviewed by [1]). Copy number variation can arise both meiotically and somatically [61] and is the primary mode by which individuals accumulate mutations. The mutation rate for CNV is 1.7 x 10-5 per locus [62] compared with 1.8 x 10-8 for SNP [63]. The median size of CNV in humans is 2.9 kb [64] and it is estimated that a typical genome has ~160 copy number variants [65] encompassing 5 to 24 Mb of the genome [66]. The distribution of CNV is nonrandom and strongly correlated with exons, segmental duplications, and transposable elements [67]. Mechanisms known to cause CNV are non-allelic homologous recombination, non-homologous end-joining, the break-fusion-bridge cycle, and replication errors following template-switching or fork stalling (reviewed by [12]). In humans and mice, CNV have been shown to account for 18 to 74% of the genetic variation in gene expression, dependent on the tissue, and at least part of the differential expression is attributable to gene dosage effects [68, 69]. Large multi-allelic CNV, which are a small subset of all CNV, cause 88% of the variation in human gene dosage [70].

Because CNV can impact phenotypic expression [65], there has been considerable interest in systematically identifying CNV in cattle to incorporate variation attributable to CNV into selection programs. Matukumalli et al. [71] used the Illumina BovineSNP50 assay to detect 79 homozygous deletions in samples from diverse *Bos taurus*, *Bos indicus* and composite breeds previously characterized by the Bovine Hapmap Consortium [72]. Bae et al. [73] also used the BovineSNP50 chip and found 368 CNVR with median size of 171 kb, and in a survey of 2,654 bulls from Italian beef and dairy breeds, Cicconardi et al. [74] identified 326 CNVR with about 60% of the CNV overlapping those identified by [73]. Similar to humans, the distribution of CNV in cattle is nonrandom with 20 to 25% of bovine CNV in segmental duplications [75, 76]. However, Hou et al. [76] showed that some SNP in CNVR have been depleted from the commercial assays, so SNP-based detection approaches are expected to underestimate the true number of CNV in the cattle genome.

Array comparative genomic hybridization (CGH) is another approach to detect CNV. Liu et al. [77] used Roche NimbleGen 385,000 probe whole-genome CGH arrays to discover 177 high-confidence CNVR in 17 breeds of cattle. Using the same assay, Kijas et al. [78] identified CNVR spanning 0.45% of the genome with a minimum detectable length of 80 kb in a small sample of Angus, Brahman and composite cattle. To increase the resolution of CNV detection, Fadista et al. [75] designed a 6.3 million probe NimbleGen CGH array, and CNVR detected ranged from 1.7kb to 2 Mb (median 16.7 kb), and spanned 23 Mb of the bovine genome. However, in 2012 Roche

discontinued manufacturing arrays and these assays are no longer commercially available.

In most of these earlier cattle studies, only large CNVR were found, probably reflecting the low resolution of these probe-based methods. Discovery of CNV in bovine whole genome sequences (WGS) from Angus, Chikso, Hanwoo, Holstein, Jeju Heugu, and Nellore bulls, has been described recently [25, 27, 79-81]. The WGS in these studies ranged from low coverage (4x) for the Holstein [79] and Jeju Heugu [80], to moderate coverage (13-22x). In most of these studies, CNV were detected based on read depth using either CNV-seq [30] or mrFAST and mrsFAST [82, 83]. The hypothesis for read depth methods is that the depth of coverage of a genomic region is correlated with copy number of the region [84].

In this study, our objective was to discover bovine CNV using WGS of the Nellore and Angus founders of our mapping population. To compare our results to previous studies, we used CNV-seq [30] for this work. We explored the influence of window size, breed (Nellore vs. Angus) of the control animal, and the depth of coverage of WGS for the control animal on the sensitivity and specificity of detection of CNVR.

II.2 Methods

II.2.1 Whole genome sequencing and alignment

All procedures involving animals were approved by the Texas A&M Institutional Animal Care and Use Committee (2011-291). There were seven Nellore (*Bos taurus indicus*) bulls and six Angus (*Bos taurus taurus*) cows, which were founders of the McGregor Genomics beef cattle population [85] that contributed to at least 10 calves in

the second generation of the cross, and from which we extracted high quality DNA from white blood cells or semen by standard proteinase K digestion and organic extraction methods. Fast Track DNA Sequencing Services (Illumina, Inc., San Diego, CA) prepared libraries for 100bp paired-end sequencing. Each animal was sequenced to a depth of at least 30x genome coverage (i.e. ~80 Gb DNA sequence) to facilitate characterization of structural variation using 2 to 6 lanes of a flowcell on a HiSeq2000 without indexing.

Raw reads were obtained from Illumina in standard fastq format [86]. After QC with fastq-mcf [87], reads were aligned to the UMD3.1 bovine assembly [88] with BWA [89], and local realignment and recalibration of quality scores was done using GATK 3.2 [90]. Bam files of these sequences are available in the NCBI short read archive [Accession numbers to be added].

II.2.2 Identification of copy number variant regions

CNV-seq [30] was used to identify CNVR based on differences in read depth after normalization for depth of coverage across the genome. The CNV-seq package consists of a Perl script (CNV-seq.pl) and an R script (cnv.R). Input best-hit files listing the chromosome and UMD3.1 coordinate were generated from the bam alignment files using samtools view [91]. For a pair of genomes (control and test), the Perl script calculates the theoretical minimum window size to obtain the best possible resolution for a desired significance level and $\log_2$ ratio, and allows the user to apply a linear scaling factor (--bigger-window) to alter the specificity of detection. Alternatively, a user-defined window size (--window-size) can be declared. The control and test genomes are

then divided into overlapping sliding windows, and the number of sequences in each window is counted (Appendix A, Figure A-1). Counts are processed as normalized ratios by the R package and specificity of CNV detection can be further refined at this point by declaring the minimum number of consecutive windows (--minimum-windows-required) deviating significantly from the $\log_2$ detection threshold required to annotate a CNV.

Initially, the Angus cow that had the highest coverage (75x) and was most unrelated to the other animals ($\widehat{\Pi} < 0.15$) was chosen as the control (A_ref) for all pairwise comparisons. Later, the influence of the breed (Nellore vs. Angus) of the control animal and the depth of coverage of the control on detection of CNVR was explored. To enable comparison to previous studies [27, 92] that detected bovine CNV using CNV-seq, we applied the same threshold values (P = 0.001 and $\log_2$ threshold = 0.7) for all pairwise tests. In comparison to the control, a doubling of the number of copies in the test genome is equivalent to a $\log_2$ ratio of 1, half the number of copies is represented by $\log_2$ ratio of -1, and an unchanged number of copies is indicated by $\log_2$ ratio of 0.

First, for each pair of WGS, we let CNV-seq calculate the theoretical minimum window size in base pairs for each chromosome, increased this window by 5, and required 10 consecutive overlapping sliding windows to be significantly different from the threshold to annotate a CNV as in [27, 92]. By this approach a slightly different window size is used for every chromosome and each pair of animals. To better evaluate the sensitivity of CNV detection and allow comparisons among the Angus and Nellore samples for windows of the same size and spacing, next we chose 7 combinations of

user-defined overlapping sliding windows and consecutive windows for annotation: 1 kb window size and 4 consecutive windows; 1 kb and 10 consecutive windows; 2 kb window size and 10 consecutive windows; 5 kb window size and 4 consecutive windows; 3 kb window size and 10 consecutive windows; 4 kb window size and 10 consecutive windows; 5 kb window and 10 consecutive windows. For these approaches the minimum detectable CNVR were 2kb, 5kb, 10kb, 10 kb, 15 kb, 20 kb, and 25 kb, respectively. For each chromosome we summarized the size of CNVR (min, max, median, mean), number of CNVR, nucleotides of CNVR, and proportion of each chromosome comprised of CNVR.

After identifying CNVR by pairwise comparisons, we investigated how many of the CNVR were detected in common among the Angus and Nellore. Because the median size (and mean size) of most of the CNVR was below 50kb, we divided the genome into windows of this size and counted how many animals had a CNVR in each window. Note that if there were two CNV with different coordinates in the same window they were treated as different CNVR, but if a CNVR was longer than 50 kb, it was arbitrarily split by the window and counted twice. Log_2 ratios from CNV-Seq for all animals simultaneously were plotted by UMD3.1 chromosomal coordinates. Positive $\log_2$ values represent insertions (gains) relative to the control animal, whereas negative $\log_2$ values represent deletions (losses).

II.2.3 Gene ontology enrichment analysis

Custom Perl scripts were written to identify the set of RefSeq genes that overlapped with CNVR on an animal-by-animal basis. DAVID [93] was used for gene

ontology (GO) enrichment analysis with default settings. Within category Benjamini-Hochberg correction was applied to control the false discovery rate.

II.2.4 Validation by quantitative PCR

Twenty CNVR (expected to be 19 gains and 1 losses) that overlapped genes, and ten CNVR that were detected by the three control animals (expected to be 4 gains and 6 losses) were chosen for validation by quantitative PCR (qPCR). As in prior studies [27, 79], we used basic transcription factor 3 (BTF3) for normalization, because it is assumed to be a single copy gene (i.e. it has one location on a chromosome). Primers (Appendix C, Additional File C-1) were designed with Primer3Plus [94] using RepeatMasked sequence extracted from the UMD3.1 assembly. Amplicon length was set to 50-250 bp and the GC clamp was set to 2. We used SYBR green chemistry in triplicate 20 µl reactions for qPCR with amplification on an ABI 7900 HT thermocycler. The same Angus cow used as the initial control animal for CNV-seq was used as the calibrator for qPCR. Data values that are not reliable were deleted. Relative copy number for each region was calculated as $2 * 2^{-\Delta\Delta C_t}$ [95, 96]. The number of copies of a CNVR in a test animal is relative to the number of copies in the control animal, but a value of 1.0 should not be interpreted as being a single copy locus because the absolute number of copies in the control animal is not known.

II.3 Results and discussion

II.3.1 Identification of copy number variants

One of the arguments for switching from chip- to sequence-based CNV detection approaches is better resolution [92]. When the window-size was calculated by CNV-seq,

the average minimum detectable CNVR (i.e. (window-size * 5) and 10 consecutive windows) was 4,953 bp, which is substantially smaller than bovine CNV detected by previous probe-based methods [73, 75]. The number of CNVR detected ranged from 1,079 to 30,603, covering 13 Mb to 163 Mb of the genome (Appendix B, Table B-1). Genome coverage of our whole genome sequences ranged from 33x to 88x per animal. We found that the number of detected CNVR was strongly correlated ($r = 0.73$; $P < 0.01$) with the depth of coverage of the tested WGS (Appendix B, Table B-2). Although the impact of a difference in the depth of coverage of the control genome compared to the test genome has been described for FREEC [97], we believe this is the first time it's been shown for CNV-seq. The total length of CNVR was comparable to that found by Janevski et al. [97] for 8 human genomes. In previous bovine studies using the same criteria for CNV-seq, the most CNV detected was 1,881 covering 30.8 Mb from WGS of ~29x coverage of the UMD3.1 reference assembly [27, 80].

We also evaluated the number of CNVR detected when we manually set the window size. A window size of 1 kb with 10 consecutive windows has a minimum detectable CNVR of 5 kb, comparable to the average minimum CNVR from our first analysis. Although the CNVR counts for these two approaches were similar for Angus, there were large differences for some of the Nellore (Appendix B, Table B-1). For example, for Nellore sample N03, only 25% of the CNVR genome coordinates detected by these two methods overlapped, whereas for the other samples 69% of the CNVR overlapped. However, it is not clear what caused this difference. We incrementally increased the size of the window to detect CNVR with minimum lengths of 2 kb to 25 kb

(Appendix B, Table B-1 and Appendix C, Additional Files C-2-C-9). For the CNVR with a minimum of 2 kb, we used 4 consecutive windows to produce the data (the default setting), but the number of detected CNVR seemed unrealistic, covering more than 10% of the length of chromosomes in most animals (Figure II.1 and Appendix B, Table B-1). We also evaluated the impact of the number of consecutive windows for CNVR with a minimum size of 10 kb, and requiring 10 consecutive windows produced more conservative counts than 4 consecutive windows, and so we used 10 windows for all other comparisons. There was a sharp decrease in the number of detected CNVR between 5 kb and 10 kb, and then a more linear decrease to 25 kb (Figure II.2; Appendix A, Figure A-2 and Appendix B, Table B-3). Regardless of the length of the window step, fewer than 10% of the CNVR were separated by less than a step (Appendix B, Table B-4), so the large inflation in counts does not appear to be an artifact of split CNV. Increasing the window size appears to lead to enrichment of multi-allelic CNVR (Figure II.3 and Appendix A, Figure A-3). The rest of our analyses are based on counts from the windows that produced CNVR of 25 kb minimum length. For this size range, there were distinct multi-allelic CNVR detected for Nellore and Angus (Figure II.4; Appendix A, Figure A-4 and A-5).

Because we had initially seen that depth of coverage was correlated with CNV counts, we repeated the analyses with a different Angus control animal (A03, 43x coverage). Janevski et al. [97] also showed that the population from which the control genome is selected impacts the number of detected CNVR, so we also used a Nellore control (N05, 45x coverage). It appears that to obtain the most conservative CNVR

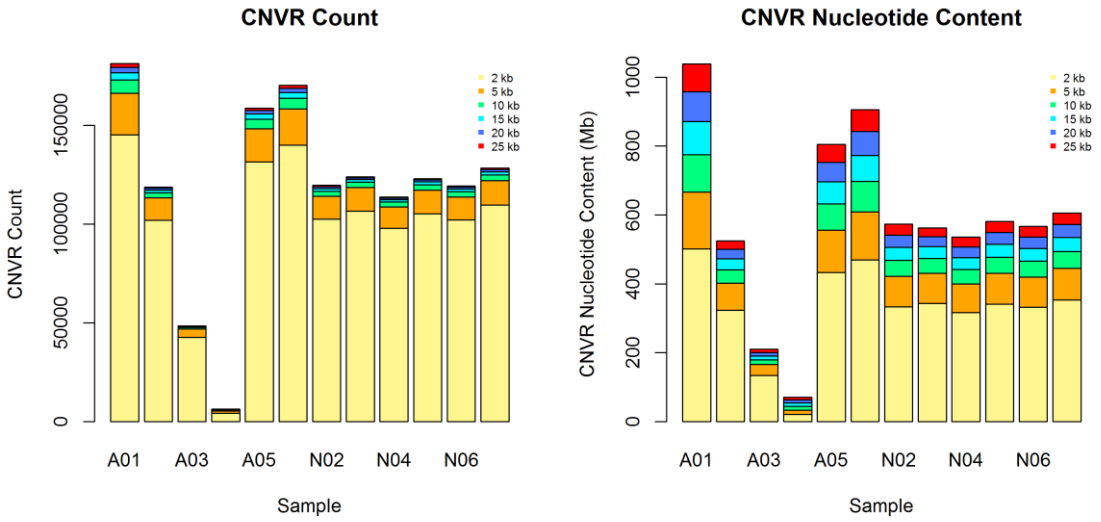


Figure II.1 Comparison of CNVR counts and nucleotide content by window size. The minimum length of detected CNVR was controlled using 10 consecutive windows and a window-size such that the minimum detectable window was half the product of the number of consecutive windows and the window size.

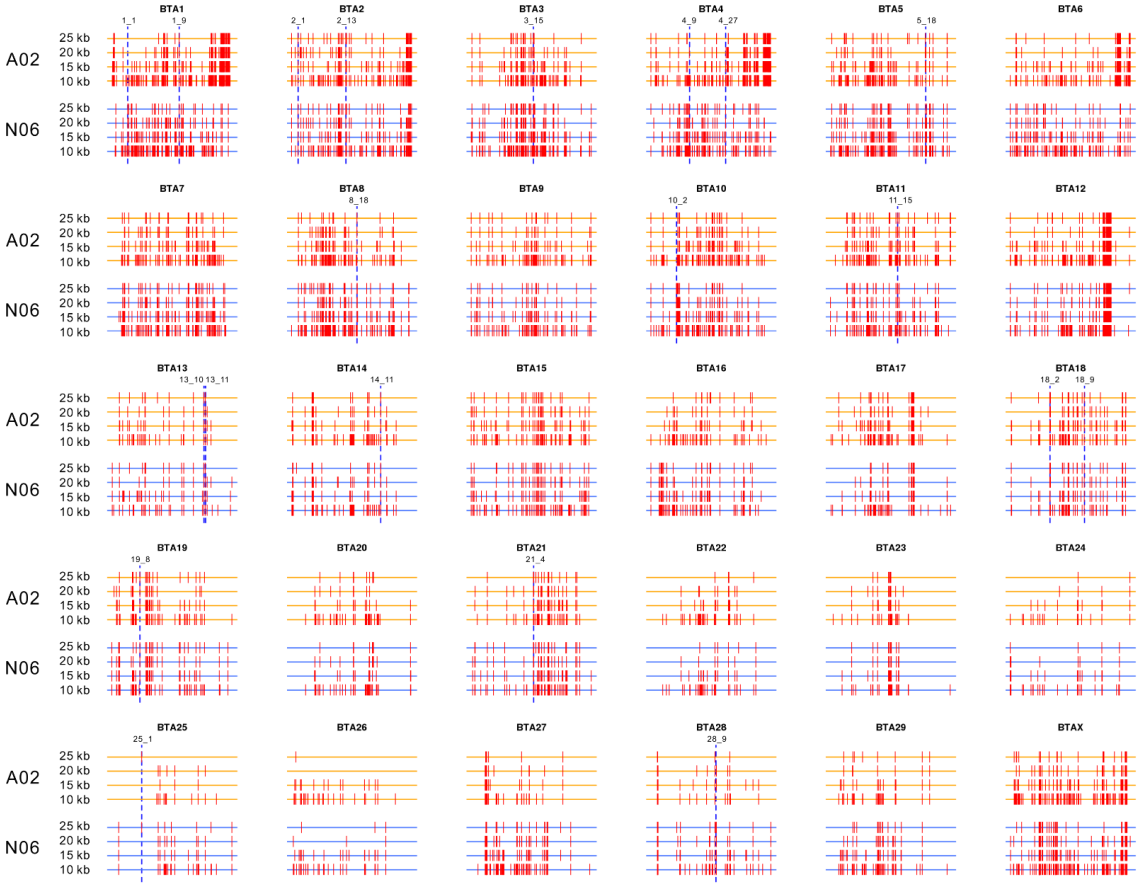


Figure II.2 Effect of window size on detection of CNVR in an Angus and a Nellore. One Nellore animal (N06; orange) and one Angus animal (A02; blue) were used as examples. Red bars indicate the positions of CNV on each chromosome and dotted blue vertical lines indicate CNV selected for validation.

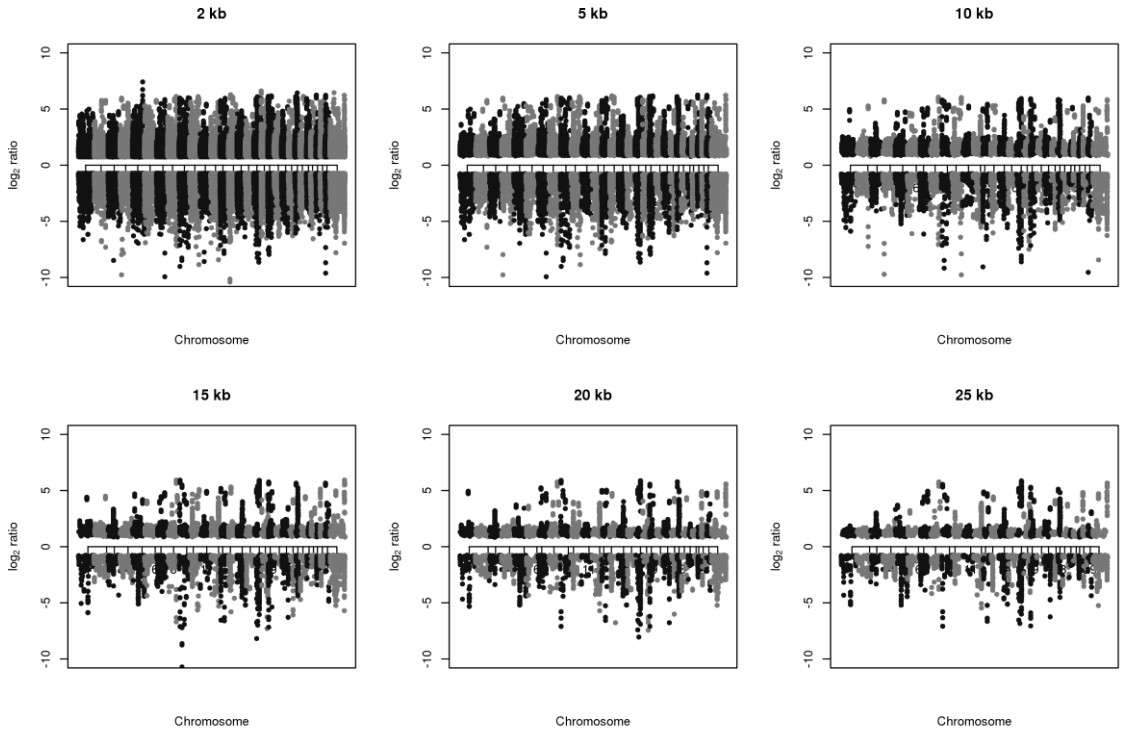


Figure II.3 Log_2 ratios for CNVR across the genome for different window sizes. X-axis represents genomic positions for chromosomes 1 to 29 and X, and the Y-axis is the log_2 ratio. In comparison to the reference, a doubling of the number of copies in the test genome is equivalent to a log_2 ratio of 1, half the number of copies is represented by log_2 ratio of -1, and an unchanged number of copies is indicated by log_2 ratio of 0. Copy number variant regions detected in any one of the animals were plotted.

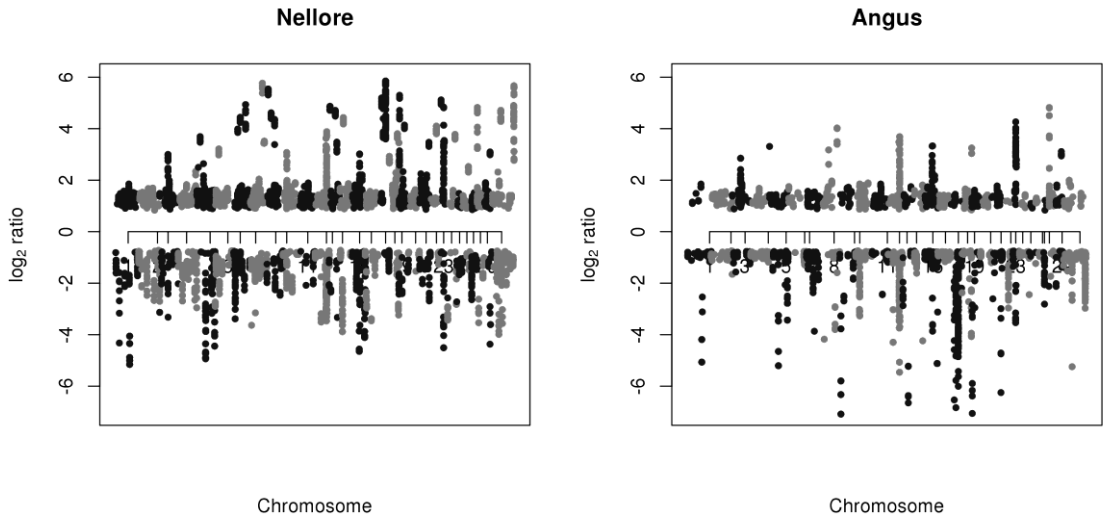


Figure II.4 CNVR identified in Nellore and Angus with a minimum length of 25 kb. X-axis represents genomic positions for chromosomes 1 to 29 and X, and the Y-axis is the $\log_2$ ratio of CNVR identified in (a) Nellore or (b) Angus. In comparison to the reference, a doubling of the number of copies in the test genome is equivalent to a $\log_2$ ratio of 1, half the number of copies is represented by $\log_2$ ratio of -1, and an unchanged number of copies is indicated by $\log_2$ ratio of 0.

counts, the control should closely match the depth of coverage of the test genome and be from the same population (Figure II.5 and Appendix C, Additional Files C-10 and C-11). Although we expected more CNV diversity between breeds than within breed, we found that more than half of the 50 kb windows containing CNVR were common to both breeds (Figure II.6). More than half of the CNVR identified by Choi et al. [27] were also identified by us with the original control animal, but there was little agreement between the CNVR we identified and those of Zhan et al. [92] (Figure II.7). There were 82 CNVR in common with Choi et al. [27] when we considered the subset found by using all three control genomes.

II.3.2 Gene ontology enrichment analysis

The subset of CNVR identified by all three control genomes was mapped to the RefSeq database because we were interested in establishing which CNVR were associated with genes. When we considered every CNVR identified in Angus, 56.1% overlapped with RefSeq compared with 60.5% for Nellore (Table II.1). Lists of CNVR mapped to RefSeq genes for each animal are in Appendix C, Additional File C-12.

When we considered CNVR from the analysis with the original reference sequence, there was evidence of enrichment for only 8 gene ontology (GO) terms in Nellore, compared with enrichment of 55 GO terms in Angus. We suspected that this large difference was a function of aligning both breeds to a *Bos taurus taurus* assembly, and thereby potentially inflating the false positive rate in Nellore and diluting any true enrichment. Indeed, for the subset of CNVR identified by all three control genomes, there were 50 enriched terms in Nellore and 40 enriched terms in Angus, and 27 terms

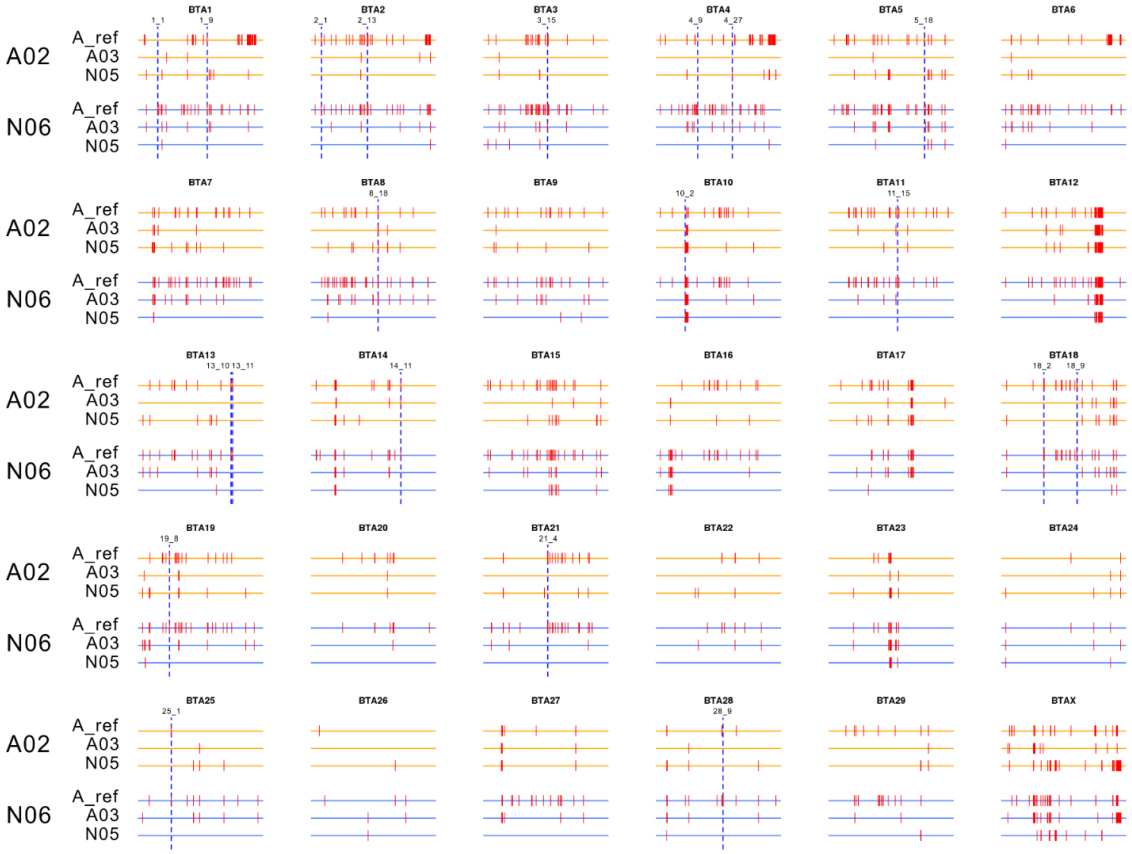


Figure II.5 Impact of different control animals on identification of CNVR. One Nellore animal (N06; orange) and one Angus animal (A02; blue) were used as examples. Red bars indicate the positions of CNV on each chromosome and dotted blue vertical lines indicate CNVR selected for validation. Each animal was compared to three different control animals: The original Angus reference (A_ref), a different Angus (A03), and a Nellore (N05).

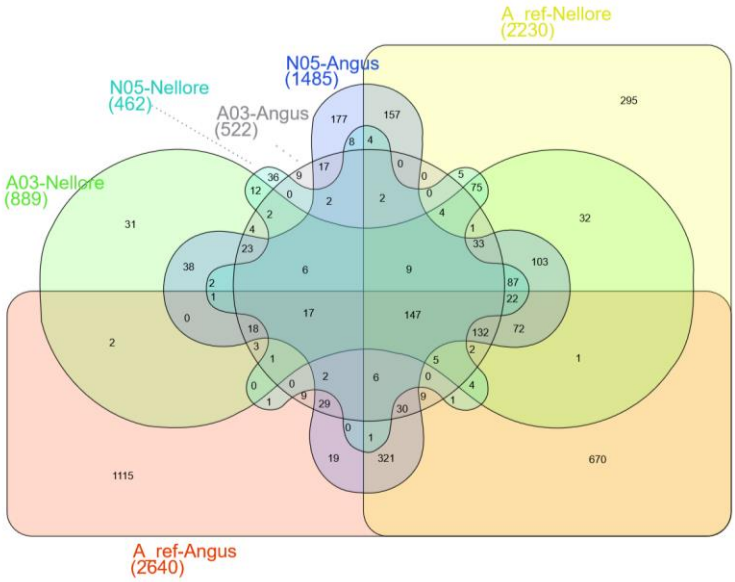


Figure II.6 CNVR detected in common among Angus and Nellore. The autosomes were divided into 50 kb windows and the number of animals with a CNVR in each window was counted. If there were two CNV with different coordinates in the same window they were treated as different CNVR. The diagram was generated by InteractiVenn by Heberle et al [98].

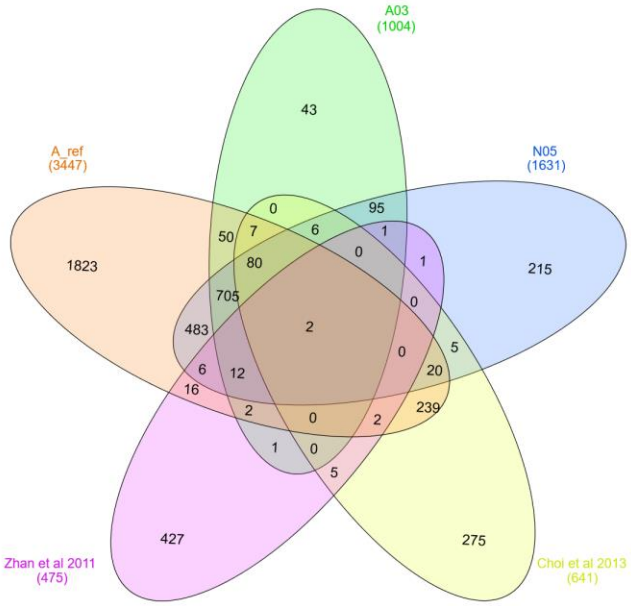


Figure II.7 CNVR identified in Nellore and Angus by CNV-seq and in common with previous studies. CNVR identified by CNV-seq in this study, Choi et al. [27], and Zhan et al. [92] CNVR were compared. The autosomes were divided into 50 kb windows and the number of animals with a CNVR in each window was counted. If there were two CNV with different coordinates in the same window they were treated as different CNVR. The diagram was generated by InteractiVenn by Heberle et al [98].

Table II.1 Proportion of CNVR mapped to RefSeq genes.

Breed	Sample	Ratio	Total Ratio
Angus	A01	0.577	0.561
	A02	0.583	
	A03	0.546	
	A04	0.570	
	A05	0.526	
Nellore	N01	0.637	0.605
	N02	0.602	
	N03	0.593	
	N04	0.634	
	N05	0.634	
	N06	0.561	
	N07	0.582	

The CNVR considered were the set with a minimum detectable length of 25 kb identified by all three references.

were enriched in both breeds. They included functions in olfactory transduction and the immune system (Table II.2 and Appendix C, Additional file C-13). Enrichment of these terms is consistent with other studies in human [99], pigs [100], and cattle [101-103] that have found enrichment of CNV associated with olfactory receptor genes, and immune system processes [25, 27].

II.3.3 Overlap of quantitative trait loci (QTL) and copy number variants

For each test animal, the subset of CNVR identified by all three control genomes was compared to AnimalQTLdb [104] and those QTL that overlapped with one or more CNVR were recovered. For Nellore, we found 1003 CNVR out of 2124 discovered CNVR with a minimum length of 25 kb overlapped 203 cattle QTL, and for Angus there were 1058 CNVR out of 2098 CNVR that overlapped 593 cattle QTL (Figure II.8 and Appendix C, Additional file C-14). In comparison, Shin et al. [81] found 2220 QTL overlapped with 6,623 putative deleted CNV.

II.3.4 Validation

The set of CNVR selected for validation was chosen from those discovered using the original control sequence. A CNVR was considered valid if there was a significant positive correlation between the $\log_2$ ratio calculated by CNV-seq and the relative copy number from qPCR. By this approach, only 15% of CNVR from the original set were validated (Table II.3; Appendix A, Figure A-6 and Appendix C, Additional File C-15) confirming there was a very high false discovery rate when the control had much deeper coverage than the test genomes. After we had run CNV-seq with the other Angus and Nellore controls, we revisited these validation results. Both of the CNVR that were

Table II.2 GO terms that are significantly enriched in both Nellore and Angus animals.

GO terms enriched in both breeds
GO:0004984~olfactory receptor activity
GO:0007186~G-protein coupled receptor protein signaling pathway
GO:0007166~cell surface receptor linked signal transduction
GO:0016021~integral to membrane
GO:0031224~intrinsic to membrane
GO:0042611~MHC protein complex
GO:0019882~antigen processing and presentation
mhc ii
SM00407:IGc1
GO:0042613~MHC class II protein complex
IPR003006:Immunoglobulin/major histocompatibility complex, conserved site
IPR003597:Immunoglobulin C1-set
IPR007110:Immunoglobulin-like
IPR013783:Immunoglobulin-like fold
PIRAPTR-SVF001991:class II histocompatibility antigen
IPR014745:MHC class II, alpha/beta chain, N-terminal
signal peptide
SM00048:DEFSN
IPR006080:Mammalian defensin
IPR001855:Beta defensin
defensin
Defense mechanisms
GO:0005253~anion channel activity
GO:0016820~hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances
GO:0043492~ATPase activity, coupled to movement of substances
GO:0015405~P-P-bond-hydrolysis-driven transmembrane transporter activity
GO:0015399~primary active transmembrane transporter activity

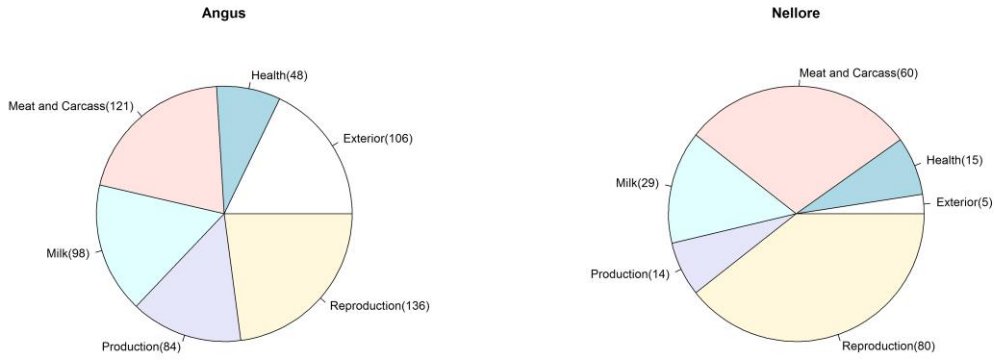


Figure II.8 Number of QTL overlapping a consensus set of CNVR. The number of QTL concordant with CNVR and detected by all comparison to all three control animals are reported by QTL category. The CNVR were identified using a 25 kb minimum window size.

Table II.3 Correlations of log2 ratios calculated by CNV-seq to Relative Copy Numbers for tested CNVR.

CNVR	Correlation	P-value	Found by all three references? (Including A04)	Found by all three references? (Excluding A04)	Validated?
1_1	-0.688	0.019158	No	No	No
1_9	-0.439	0.176507	No	No	No
2_1	-0.526	0.096323	No	No	No
2_13	0.250	0.485957	No	No	No
3_15	-0.390	0.236328	Yes	No	No
4_9	0.811	0.002439	No	No	Yes
4_27	0.805	0.00281	Yes	Yes	Yes
5_18	0.354	0.285762	Yes	No	No
8_18	0.373	0.258567	Yes	No	No
10_2	0.689	0.018929	Yes	Yes	Yes
11_15	0.167	0.624396	Yes	No	No
13_10	0.086	0.80242	No	No	No
13_11	-0.313	0.348928	Yes	No	No
14_11	-0.244	0.470441	Yes	No	No
18_2	-0.193	0.568972	Yes	No	No
18_9	-0.055	0.873244	No	No	No
19_8	-0.050	0.882528	No	No	No
21_4	-0.485	0.130627	Yes	No	No
25_1	-0.245	0.467418	No	No	No
28_9	-0.601	0.05029	No	No	No

detected using all three controls validated. One of those was the deletion CNVR 4_27, which spans 75026251-75116250 on BTA4 (Figure II.9). Nine of the ten more CNVR identified by the three control animals were validated because they have variable number of copies compared to the control A_ref, as shown in Appendix A, Figure A-7.

II.4 Conclusions

We observed that the number CNVR discovered using CNV-seq software was strongly correlated with the depth of coverage of the tested genome compared to the reference genome. The normalization algorithm of CNV-seq did not appear to adequately overcome large differences in depth of coverage and, consequently, the false discovery rate was grossly inflated. We overcame this issue by using different animals as the control and focusing on the common set of CNVR, which is summarized in Appendix C, Additional File C-16. We chose a conservative 25kb minimum window size for our final analyses because there appeared to be enrichment for multi-allelic CNVR as window size increased. Multi-allelic CNVR have previously been associated with variation in gene dosage and differential expression of genes. We showed that the CNVR discovered in Nellore and Angus covering RefSeq genes were enriched for olfactory transduction and immune system function. About half of the discovered CNVR were associated with previously identified cattle QTL for production, reproduction and health. We expect that future work with the population founded by the individuals characterized in this study will show that some of the discovered CNV contribute to variation in these important phenotypes.

Figure II.9 Log_2 ratios for putative CNVR 4_27 on BTA 4 from 75026251-75116250 bp. The log_2 ratios calculated by CNV-seq were plotted for each Angus (A) and Nellore (N). In comparison to the reference, a doubling of the number of copies in the test genome is equivalent to a log_2 ratio of 1, half the number of copies is represented by log_2 ratio of -1, and an unchanged number of copies is indicated by log_2 ratio of 0.

CHAPTER III

COMPARISON OF THREE SOFTWARE APPLICATIONS FOR COPY NUMBER VARIANT DETECTION AND THE ALGORITHMS BEHIND THEM

III.1 Introduction

Previously, cytogenetic technologies including karyotyping and fluorescence *in situ* hybridization [105], array-based comparative genomic hybridization, and single-nucleotide polymorphism array approaches [106] have been used to identify CNV. Limitations of them include hybridization noise, inability to discover novel and rare CNV, low genome coverage, and low resolution [107, 108]. In order to overcome the limitations of these traditional approaches, many WGS based methods have been developed. Currently, a variety of software applications based on various algorithms are publicly available. The major algorithms used include read depth (RD), read-pair (RP, or paired-end mapping), split read (SR), *de novo* assembly (AS), and the combined approaches based on them [109], as shown in Figure III.1.

Read depth-based methods are major methods for CNV identification in NGS data. They compare RD of the test genome to a reference genome to identify CNV. The assumption underlying them is that in a genomic region, the depth of coverage is correlated to the copy number [32]. Read depth is calculated by counting the number of mapped reads in predefined windows. Read depth is then normalized and the potential biases from GC content and repetitive regions are corrected. Finally, copy numbers are detected along the chromosome and the genomic regions with similar copy numbers are

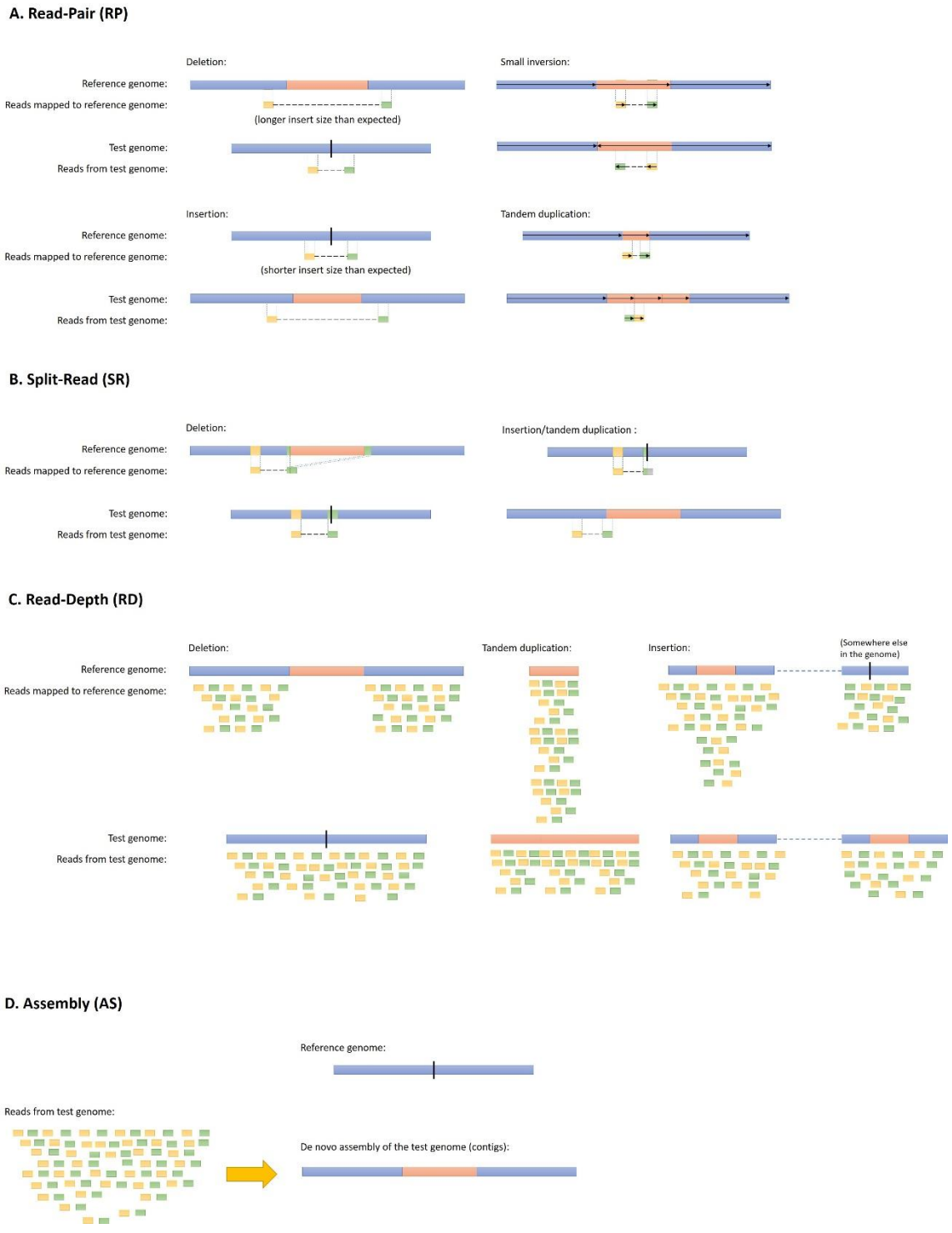


Figure III.1 The major algorithms used in CNV identification for NGS data.

merged to one CNVR [110]. Read depth-based methods can use a single sample, paired samples, or pooled samples [109]. If using a single sample, absolute copy numbers are calculated, whereas if using paired case/control samples, relative copy numbers of the test genome compared to the control genome are calculated [109]. Software applications based on RD methods include BIC-seq [111], CNV-seq [30], and CNVnator [112]. Among them, CNV-seq is designed to identify insertions (gains) and deletions (losses) in the genome compared to a control genome [30]. After normalization for depth of coverage across the genome, it detects CNV based on differences in RD. CNV-seq calculates a $\log_2$ value from which the relative copy numbers compared to the control genome can be obtained. It is widely used in CNV identification in cattle and humans.

Read pair-based methods are based on the comparison of average insert size of DNA fragments. In paired-end sequencing, the insert sizes are expected to follow a specific distribution [113]. Read pair-based methods detect the discordance of the average insert size between sequenced RP from a test genome and the expected size from the control genome [109]. There are two frequently used approaches: the clustering approach, which uses a predefined distance to detect discordant reads, and the model-based approach, which uses a probability test to find any unusual distances between read pairs compared to the distance distribution of the genome [114]. Limitations of RP-based methods include: not detecting insertions that are larger than the average insert size of the genome library [115], and not detecting CNV in regions of low-complexity with segmental duplications [114]. Software applications using RP-based methods include PEMer [116] and BreakDancer [47]. BreakDancer [47] detects a variety of structural

variants (SV) including CNV. The methods used by BreakDancer are BreakDancerMax, which utilizes the clustering approach, and BreakDancerMini, which utilizes the model-based approach [114]. BreakDancerMini detects small indels (10–100 bps) whereas BreakDancerMax detects insertions, deletions, inversions, intra-chromosomal translocations, and inter-chromosomal translocations. But we are only considering insertions and deletions as CNV in this study. One more limitation for BreakDancer is its results are not reliable in repetitive regions, because each read is only assigned to one cluster, and the reads which can be mapped to several genomic regions are discarded by BreakDancer, even if they have a high mapping quality [114].

Split read-based methods utilize reads where one read from the pair has a reliable mapping, but the other one fails to map to the genome or is only partially mapped, which becomes a potential source of the break-point for the CNV at the single base pair level [117]. The incompletely mapped reads are split into fragments, and the first and last fragments of each split read are aligned to the genomic reference sequence independently to identify the precise start and end points of the CNV [114]. The limitations of SR-based methods are that they heavily rely on read length and only work with the unique regions of the reference genome [114]. Applications based on SR methods include Pindel [118] and AGE [119].

Another approach is *de novo* assembly, which assembles a new genome based on short reads and compares it to the original genomic reference sequence to identify genomic regions with CNV [114]. This method is unbiased because it does not rely on read alignment, but it also has limitations: a minimum read coverage is required but high

coverage increases the complexity of assembly; the quality of the assembled contigs are low for non-human organisms; and there is high demand for computational resources [114]. Applications based on assembly include Magnolya [120] and the Cortex assembler [121]. Due to the limitations of this method, it was not used or compared in this study.

Among the above methods, RD-based methods can detect the exact copy number of CNV, whereas RP and SR based methods only detects the position of CNV. Also, RD-based methods can identify large insertions and CNV in complex genomic regions, which are difficult for RP- and SP-based methods [53].

Because each method has their own strengths and limitations, combined approaches are created to take advantage of the best features of the various methods. Step-wise approaches are used to combine data from two or more sources [109]. Applications like this are CNVer [46], which utilizes both RP and RD information, and RAPTR-SV [52], which is based on the combination of RP and SR methods to detect SV, and does not require a control genome. RAPTR-SV first identifies discordant RP, and places overlapping discordant RP with the same orientation into the same groups. Finally, SR are assigned by searching for complete and one-end unmapped alignments of split pairs, and CNV are identified [52].

In this study, we compare the performance of three software applications: BreakDancer, CNV-seq, and RAPTR-SV to identify CNV regions (CNVR) in the Nellore and Angus founders of a beef cattle mapping population, and discuss the advantages and shortcomings of the algorithms behind them. We also propose a way to

obtain a more reliable set of CNVR with fewer artifacts and noise by finding the CNVR set that are concordant across all three algorithms.

III.2 Methods

III.2.1 Aligned sequence data

Bam files of seven Nellore (*Bos taurus indicus*) bulls and six Angus (*Bos taurus taurus*) cows, which were founders of the McGregor Genomics beef cattle population [85] (see II.2.1) were used for this study.

III.2.2 Identification of copy number variant regions

Three software applications were used to identify CNVR: CNV-seq [30], BreakDancer [48], and RAPT-SV [52].

To enable comparison to CNV studies in Chapter II and previous bovine CNV studies by CNV-seq [27], the same strict threshold values ($P = 0.001$ and $\log_2$ threshold = 0.7) and a minimum detectable CNVR size of 25 kb (window-size = 5000, consecutive-window = 10) were used. These parameters were used because a 25 kb minimum detectable CNVR size was demonstrated to be a suitable window size for CNV detection for these sequence data in Chapter II. And the control animal for all pairwise comparisons was the same Angus cow (A_ref) as in Chapter II. This Angus cow had the highest coverage (75x) and was the most unrelated to the other animals ($\widehat{\Pi} < 0.15$).

Default settings of BreakDancer were applied for detection of structural variants. The same Angus animal (A_ref) was used as the control. Default setting of RAPTR-SV were used as well. No control animal was needed for this software application. After

identifying structural variants by RAPTR-SV, a filtration step was applied following the instructions on Git-hub for RAPTR-SV.

Because RAPTR-SV categorizes insertions into two categories: insertions (one copy inserted somewhere else in the genome) and tandem duplications (one or multiple copies inserted next to the original sequence), we categorized insertions in the same way. The results from RAPTR-SV and BreakDancer were first filtered to omit other structural variants and to include CNVR larger than 1 kb, and then for comparison with CNV-seq they filtered to the size of 25 kb to 502.5 kb, which was the minimum and maximum sizes of CNVR detected by CNV-seq.

After identification of CNVR by the three software applications, we investigated how many of the CNVR were detected in common among the applications and breeds (Nellore and Angus). The genome was divided into 50 kb nonoverlapping windows because the median (and mean) sizes of all CNVR detected were close to 50 kb. The number of animals having a CNVR in each window was then counted. Like in II.2.2, if two CNV with different coordinates appeared in the same window, they were treated as two different CNVR. For CNVR longer than 50 kb, they were arbitrarily split by the window and counted twice.

The performance of the three software applications was compared using R and custom Perl scripts. Venn diagrams were generated by InteractiVenn [98].

III.2.3 Validation

Ten CNVR identified by the three applications are chosen for validation using the same method as in II.2.4. Primers are summarized in Appendix C, Additional File C-1.

III.3 Results and discussion

III.3.1 Identification of copy number variant regions

Results from BreakDancer, CNV-seq and RAPTR-SV are summarized in Appendix B, Table B-5 and Appendix C, Additional File C-9 and Additional File C-17. Compared to CNV-seq, BreakDancer and RAPTR-SV detected much more insertions, deletions and tandem duplications for each of the animals before filtration, which included both small indels and large CNVR. The average number of CNVR detected for each animal was 19201 for BreakDancer, 809 for CNV-seq and 23149 for RAPTR-SV. There were on average 1,864 and 456 CNVR for BreakDancer and RAPTR-SV, respectively, after filtering the CNVR length to 25 kb to 502.5 kb. Examination of the distribution of CNVR identified by CNV-seq, shows that it identifies insertions, including tandem duplications, and deletions across the genome. BreakDancer and RAPTR-SV detect more SV types across chromosomes and unmapped contigs, but only insertions, tandem duplications, and deletions were retained for further analysis. Before filtration, BreakDancer and RAPTR-SV detected large proportions of small variants compared to CNV-seq (Appendix B, Table B-5): the mean and median sizes for BreakDancer were 1068.97 kb and 0.68 kb; the mean and median sizes for RAPTR-SV

were 9.79 kb and 0.17 kb; the mean and median sizes for CNV-seq: 45.90 and 32.71 kb. Note that the mean size for BreakDancer were large because it detected some extremely large CNVR, but the majority of CNVR detected by BreakDancer were very small. After filtering the CNVR size to larger than 1 kb, which is the minimum size of a CNV by definition, insertions and tandem duplications for BreakDancer disappeared completely, and insertions for RAPTR-SV disappeared as well, which confirmed that RP-based methods cannot detect insertions larger than the average insert size of the genome library [115], and SR based methods heavily rely on read length [114].

The remaining types of CNV identified by the applications after filtration are shown in Table III.1. The type and number of CNVR identified by the three applications in one Angus animal and one Nellore animal are also shown as an example in Figure III.2. After filtration, only 9.7% and 2.1 % of the total CNVR remained for BreakDancer and RAPTR-SV, respectively.

The number, mean, and median sizes of CNVR identified by the three applications after filtration for each animal are summarized in Table III.2 and Figure III.3. The mean and median sizes for BreakDancer were 119.06 kb and 76.20 kb; the mean and median sizes for RAPTR-SV were 138.81 kb and 90.11 kb. Both the mean and median sizes for BreakDancer and RAPTR-SV were higher compared to those of CNV-seq after filtration. There were large differences in the type and number of CNVR identified by the three applications. RAPT-SV detected much fewer CNVR than BreakDancer and CNV-seq. BreakDancer detected the most CNVR. The number of CNVR detected by CNV-seq was not consistent among the animals, whereas it was

Table III.1 Types of structural variants identified by BreakDancer, CNV-seq and RAPTR-SV.

Name	Method	Need Reference?	All SV types detected	CNV types detected after filtration of range
BreakDancer	BreakDancerMax and BreakDancerMini (based on read pair)	Yes	Insertions (including tandem duplications), deletions, inversions, intra-chromosomal translocations, inter-chromosomal translocations	Deletions
CNV-seq	Read depth	Yes	Insertions (including tandem duplications), deletions	Insertions (including tandem duplications), deletions
RAPTR-SV	Combination of read pair and split-read	No	Insertions, deletions, tandem duplications	Deletions, tandem duplications

The "All types" shows all SV identified by each of the applications, and the "CNV types" shows all CNV types identified by the applications after filtration by length.

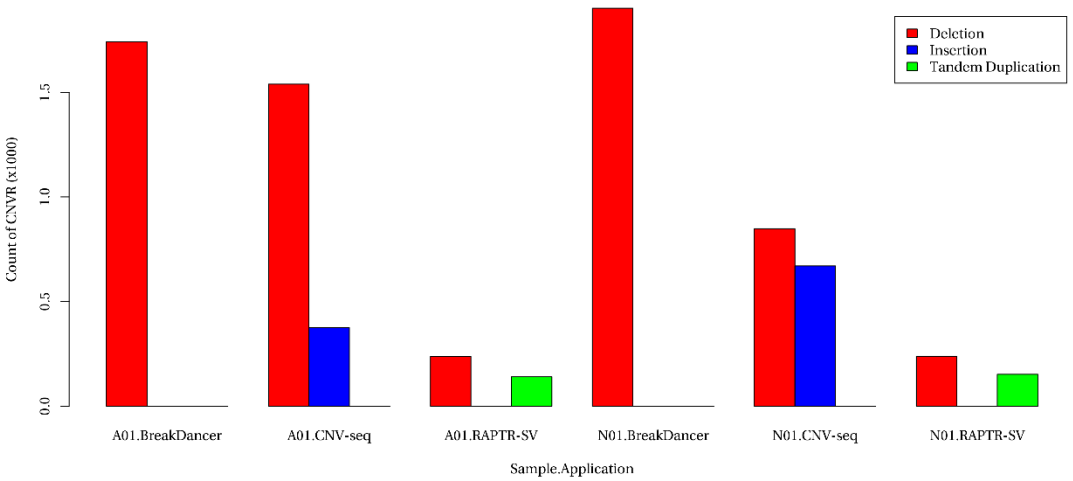


Figure III.2 Type and number of CNVR identified by BreakDancer, CNV-seq and RAPTR-SV. One Nellore animal and one Angus animal are shown as an example.

Table III.2 Comparison of lengths of CNVR identified by BreakDancer, CNV-seq and RAPTR-SV.

Sample	Number			Mean (kb)			Median (kb)		
	Break-Dancer	**CNV-seq**	**RAPTR-SV**	**Break-Dancer**	**CNV-seq**	**RAPTR-SV**	**Break-Dancer**	**CNV-seq**	**RAPTR-SV**
A01	1742	1914	377	119.36	42.16	146.53	77.07	32.5	94.86
A02	1785	628	447	116.86	39.73	135.7	72.97	30	85.87
A03	1945	165	465	118.12	54	137.87	76.7	40	89
A04	1931	141	475	117.48	59.27	131.83	76.42	37.5	88.96
A05	1942	1271	365	119.07	41.24	144.97	77.47	32.5	96.59
N01	1902	1518	390	120.91	42.24	141.25	77.9	32.5	93.48
N02	1946	679	473	120.8	46.97	138.04	77.46	30	89.83
N03	1972	594	536	119.61	44.57	138.96	75.48	30	89.72
N04	1769	603	510	116.77	47.64	142.31	72.66	32.5	89.07
N05	1717	712	441	118.67	44.95	136.56	75	32.5	89.09
N06	1738	697	501	119.39	44.62	136.48	77.3	30	90.17
N07	1977	790	479	121.7	43.45	141.71	77.97	32.5	89.39
A_ref	-	-	468	-	-	132.37	-	-	85.43

The number of CNVR, mean and median sizes of CNVR identified by the three applications after filtration are summarized. BreakDancer identified the most CNVR counts. CNV-seq detected the lowest mean and median CNVR sizes while RAPTR-SV detected the highest mean and median CNVR sizes. Because RAPTR-SV does not need a control animal, it gives the information of A_ref as well.

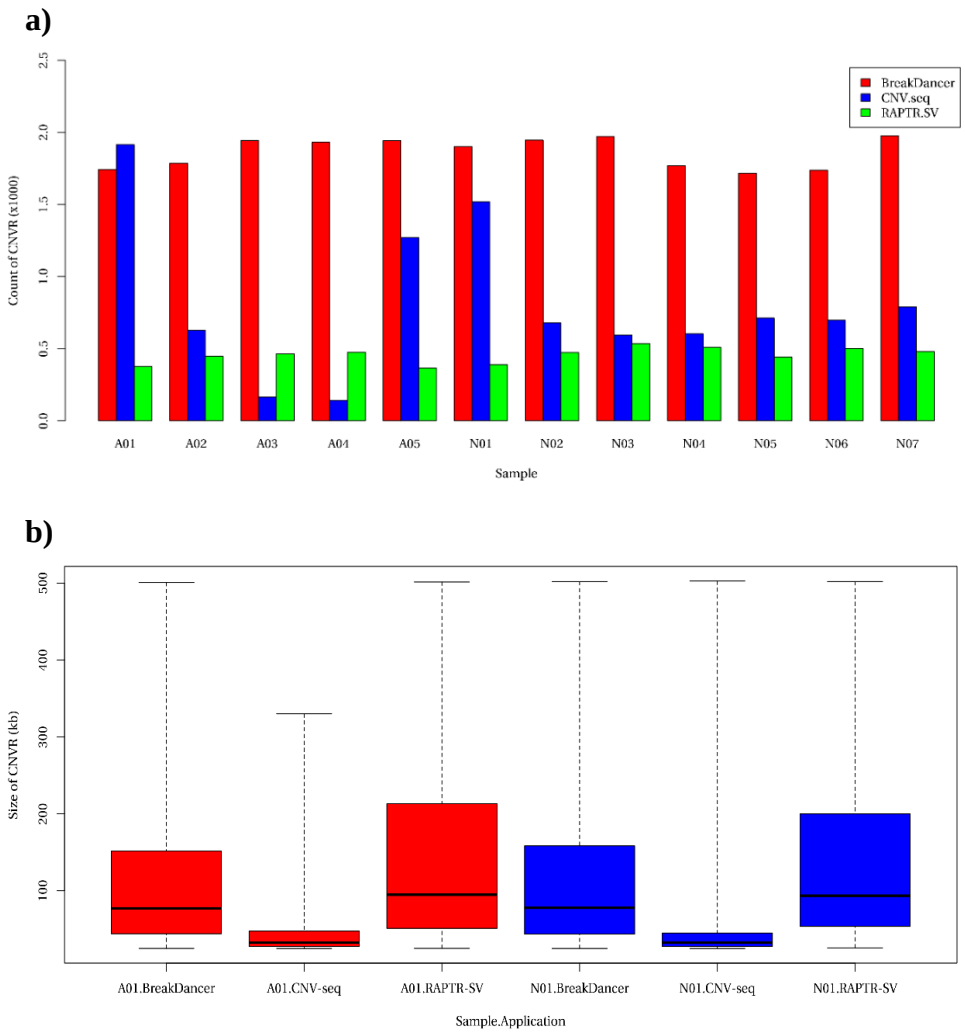


Figure III.3 Comparison of number and size for CNVR identified by BreakDancer, CNV-seq and RAPTR-SV. a) The number of CNVR identified by the three applications are summarized after filtration. b) Boxplot of CNVR sizes identified by BreakDancer, CNV-seq and RAPTR-SV in one Angus animal and one Nellore animal. Size of CNVR are after filtration. CNV-seq tend to identify smaller CNVR than the other two.

consistent for the other two applications, which may indicate that the RD-based method is more sensitive to depth of coverage of the sequences. Although CNVR detected by the three applications were filtered to the same size range, CNV-seq had the lowest mean and median CNVR sizes, therefore it detected the smallest CNV after filtration. RAPTR-SV identified CNVR with the highest mean and median sizes, which were slightly higher than the mean and median of BreakDancer. RAPTR-SV was the only application among the three that did not need a control animal, so it provided information on the control animal, A_ref, as well.

As shown in Table III.2 and Figure III.3 (a), the number of CNVR identified by CNV-seq had great differences among Angus animals, but were more consistent among Nellore animals. Because the control animal was an Angus, this further demonstrates the point in Chapter II that for CNV-seq and RD-based methods, CNVR counts are related to depth of coverage of the test genome compared to the control genome. The number of CNVR detected by BreakDancer and RAPTR-SV were more consistent among all Nellore and Angus animals compared to CNV-seq. This shows that the performance of RD methods highly rely on depth of coverage of the test genome compared to the control genome, whereas RP and SR methods do a better job in handling different depth of coverage among different samples.

Copy number variant regions identified were then classified in the consecutive 50kb windows on each of the chromosomes, and the distribution of CNVR by BreakDancer, CNV-seq and RAPTR-SV across the genome was analyzed. One Nellore and one Angus animal are shown in Figure III.4 as examples. CNVR identified by

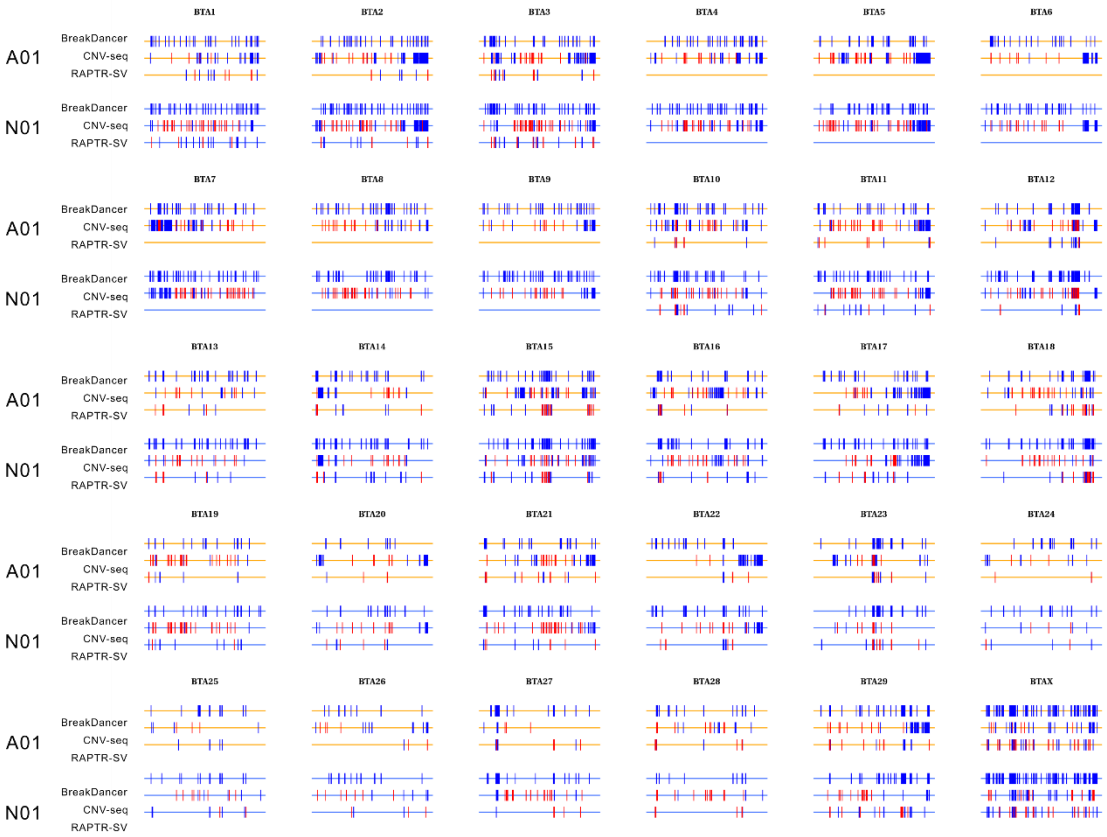


Figure III.4 CNVR distribution by BreakDancer, CNV-seq and RAPTR-SV across the genome. CNVR identified by BreakDancer, CNV-seq and RAPTR-SV on each of the chromosomes of an Angus animal and a Nellore animal are shown. CNVR were classified in the consecutive 50 kb windows on each of the autosomes, and CNVR sizes are after filtration. Blue indicate deletions and red indicate insertions and tandem duplications.

RAPTR-SV were distributed across the genome with low density, while CNVR detected by BreakDancer and CNV-seq showed a relatively high density distributed uniformly across the chromosome. The three software applications identified some common CNVR, which indicates that those CNVR are more likely to be authentic structural differences biologically rather than artificially generated from the sequences. However, we notice that some common CNVR were identified as insertions or tandem duplications according to one or two applications, but were identified as deletions according to the other application(s) (Figure III.4), which indicated discordance among different applications and the methods behind them. For example, the common set of CNVR identified by the three applications include insertions, but no insertions were detected by BreakDancer.

The agouti signaling protein (ASIP) gene, which locates on BTA13:64213312 – 64239964, is known to have CNV in exon regions in ovine genome that affect coat color [122, 123]. ASIP gene in the ovine genome has similar organization of the exon-introns compared to that in the bovine genome, and mutations in this gene cause “non-agouti” mutations in livestock [122]. The region of 60 Mb to 70 Mb on BTA13 was extracted to compare CNVR identified by the three software applications in the ASIP gene. One Nellore animal (N01) was shown in Figure III.5 as an example.. Only two insertions were detected in this region for CNV-seq with 25 kb minimum detectable CNVR size, whereas RAPTR-SV and BreakDancer both detected deletions. These deletions had no overlap with the two insertions. Interestingly, the deletion detected by RAPTR-SV overlapped with one deletion detected by BreakDancer, which indicates that the

a)

BreakDancer

CNVR	Chromosome	Position 1	Position 2	Type	Size
1	13	61639838	61716178	DELETION	76371
2	13	63036192	63158438	DELETION	122228
3	13	63042153	63165534	DELETION	123334
4	13	66423045	66450620	DELETION	27537

CNV-seq with 25 kb minimum detectable CNVR size

CNVR	Chromosome	Start	End	Type	Size
1	13	64436251	64463750	INSERTION	27500
2	13	65566251	65591250	INSERTION	25000

RAPTR-SV

CNVR	Chromosome	Start	End	Type	Size
1	13	61639843	61716187	DELETION	76345

b)

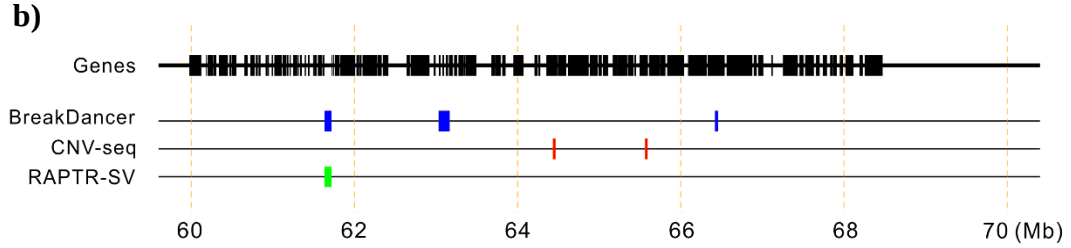


Figure III.5 CNVR identified by BreakDancer, CNV-seq and RAPTR-SV between 60 Mb to 70 Mb on BTA13 in Nellore N01. Chromosome, start, end, type and size of each CNVR detected in this region are listed in a). The genes and CNVR in this region are plotted in b). Blue and green indicate deletions, whereas red indicate insertions.

outcomes of RP method are relatively consistent in different software applications, and SR method is more consistent with RP than RD method. However, this deletion doesn't overlap with interesting genes, and none of the CNVR detected overlapped with ASIP gene.

The CNVR shared by BreakDancer, CNV-seq with 25 kb minimum detectable CNVR size, and RAPTR-SV for the set of Angus and Nellore animals are shown in Figure III.6. Large proportions of CNVR common to the three applications were shared by CNV-seq and RAPTR-SV. Distribution of CNVR shared by all three software applications in Nellore and Angus breeds are shown in Table III.3. For all of the applications, more CNVR were shared between the two breeds than appeared in one breed only. The proportion of CNVR shared between breeds were high (about 61.7% ~ 75%), which was consistent with the results from CNV-seq using various minimum detectable CNVR sizes in II.3.

III.3.2 Comparison of performance and algorithms behind the three software applications

Of the three applications used in this study, CNV-seq is the only one that generates $\log_2$ of the copy number ratio, which makes it possible to calculate the relative number of copies and copy number change for a CNVR. It also detects more insertions than deletions, in contrast to BreakDancer and RAPTR-SV. CNV-seq has several parameter settings to change, which greatly affect the number and size of CNVR detected, whereas the other two packages don't have these parameters to adjust. BreakDancer and CNV-seq both require a control animal, so it is important to choose the

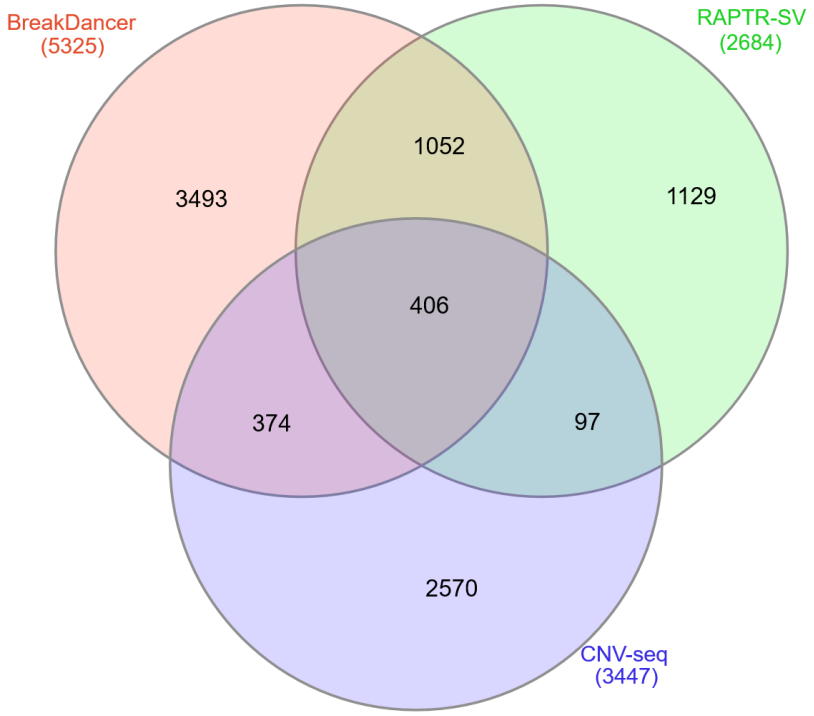


Figure III.6 Venn Diagram of CNVR shared between different software applications. CNVR identified by BreakDancer, CNV-seq with 25 kb minimum detectable CNVR sizes and RAPTR-SV for all Angus and Nellore animals are shown. CNVR were classified in the consecutive 50kb windows on each of the chromosomes and CNVR sizes are after filtration. The same windows appeared for the same software application are only calculated once. The diagram was generated by InteractiVenn by Heberle et al [98].

Table III.3 Distribution of CNVR shared by all three software applications in Nellore and Angus breeds.

	Angus Only	**Common**	**Nellore Only**
BreakDancer	18	351	37
CNV-seq	63	188	155
RAPTR-SV	41	283	82

CNVR shared by all three software applications are classified into three categories: appear in Angus only, shared between Nellore and Angus, and appear in Nellore only. CNVR were classified in the consecutive 50kb windows on each of the chromosomes and CNVR sizes are after filtration.

most suitable control animal. The selection of the control animal for CNV-seq was discussed in II.2.2. RAPTR-SV is the only application that does not require a control sequence, and is the only application we used that categorizes insertions and tandem duplications separately. When the minimum and maximum size of CNVR were determined, BreakDancer detected the most CNVR. Unlike the other two applications, the number of CNVR were not uniform among samples with different depths of coverage in CNV-seq.

Because we wanted to compare the performance of the three applications on the same scale, we filtered all the CNVR sizes detected to 25 kb to 502.5 kb. The mean, median and maximum sizes of CNVR detected by CNV-seq were relatively low compared to the other two applications after filtration. Before filtration, for CNV-seq, the minimum sizes (25 kb) were much higher and the maximum sizes were much lower (502.5 kb) compared to the other two applications (Appendix B, Table B-5). When not filtered, BreakDancer and RAPTR-SV had very low median and minimum CNVR sizes (Appendix B, Table B-5). It is reasonable because RP and SR methods have problems detecting large insertions, and BreakDancer also has a method for detecting small indels. This observation is similar to Yoon et al. [53] that showed RD-based methods had a better performance in detecting large CNV, which was not the case for RP- and SR-based methods. Pirooznia et al. [109] also pointed out that it was difficult for SR-based methods to identify large-scale SV, which is similar to our findings that the median and mean CNVR sizes of RAPTR-SV are much smaller than that of CNV-seq (Table III.2).

However, BreakDancer and RAPTR-SV did detect many CNVR with reasonable length (>1 Mb) according to our findings (Appendix B, Table B-5).

The CNVR identified by RD and RP methods differ greatly. The performance of RD-based methods highly rely on the depth of coverage of the test genome compared to the control genome, selection of the control animal, and selection of the window size. For RD-based methods, larger windows achieve higher confidence for CNV detection, but this results in more opportunities to miss small CNV. Therefore, it is important to choose suitable window sizes in order to obtain CNVR with desirable lengths. The normalization algorithm of CNV-seq did not appear to adequately overcome large differences in depth of coverage and, consequently, the false discovery rate was grossly inflated. This issue was overcome by using different animals as the control and focusing on the common set of CNVR, as shown in II.4.

RP and SR based methods are more stable compared to RD, since the CNVR discovered are not that sensitive to parameter settings compared to RD-based methods. BreakDancer detects a large quantity of small CNVR before filtration, which is contrary to Medvedev et al. [115] who found that RP-based methods are insensitive to small insertions and deletions because it is difficult for them to separate small perturbations in read-pair distance from the normal background variability. One explanation may be that the accuracy of RP-based methods highly rely on insert size of sequencing libraries. Large-insert libraries may miss small insertion and deletion events [124].

Nowadays, combined methods are more popular since any method alone has its strengths and shortcomings because of the complex underlying structure of the SV sites,

and none of them are sufficiently comprehensive [125]. However, it was reported that combined methods did not perform well against duplications or repeats [125]. RAPTR-SV, which uses a combination of read-pair and split read methods, detected a reasonable amount of CNVR including both tandem duplications and deletions, and the identified CNVR had many overlap with CNVR identified by BreakDancer and CNV-seq. Therefore, this combination of RP and SR methods is the most stable and consistent among the three, with the advantage of not needing a control animal.

Since the limitations of each method may result in false discovery or only detecting a subset of CNV, it is important to find ways to obtain useful information from outputs of various applications and utilize them all. We chose to focus on the consensus set of CNVR (Appendix C, Additional File C-18) for future studies to overcome the limitations of individual methods. However, we noticed that a proportion of CNVR in the consensus set are insertions, which should not present because all CNVR found by BreakDancer are deletions. We still chose to use the consensus set because, except for the contradiction in losses and gains, there must be structural variation in those regions due to their identification by all three applications.

III.3.3 Validation

The ten CNVR were validated because they have variable number of copies compared to the control A_ref, as shown in Appendix A, Figure A-8.

III.4 Conclusions

The performance of RD-based methods highly rely on the depth of coverage of the tested genome compared to the reference genome, selection of a control animal, and

selection of the window size. Larger windows achieve higher confidence for CNV detection, but results in higher opportunities of missing small CNV. Read pair- and RAPTR-SV-based methods are more stable compared to RD, but it is difficult for them to identify large insertions. Also, SR based methods have problems detecting large-scale SV, including CNV. Except for these limitations, the combination of RP and SR methods used by RAPTR-SV is the most stable and consistent among the three, with the advantage of not needing a control animal.

Few studies were done using multiple methods to detect SV including CNV in cattle. This study shows how these software applications and their underlying methods work and the performance of them, providing a good reference for method selection for CNV identification in future studies.

CHAPTER IV

GENOME-WIDE ASSOCIATION STUDY OF SNP-TAGGED COPY NUMBER VARIANT REGIONS IN A BEEF CATTLE MAPPING POPULATION

IV.1 Introduction

Copy Number Variants may be involved in the formation of different phenotypes in the cattle populations due to the dosage variability of the genomic sequences underlying them [126]. Although sequence-based approaches for CNV detection provide higher resolution compared to chip-based approaches [92], it is expensive to perform WGS on a large population. However, it is known that CNVR can be tagged by nearby SNP, and the tagged SNP could be used to capture genetic variation and association with phenotypes [64, 127]. Therefore, if CNVR were detected from a few animals in a large population, SNP identified in that population having high linkage disequilibrium (LD) with these CNVR could be used to assess the association of these CNVR with interesting traits and predict the effect of these traits in the population.

Genome-wide association studies are typical methods for assessing association of genetic variants, including SNP and CNVR, with traits and their effects on phenotypic expression. These studies model the effects of genetic variants genome-wide in a population to see if any variants are strongly associated with a specific trait of interest. Genome-wide association studies are often used to study diseases with complex genetics in humans [128], and they are also used in other organisms, including livestock. Although normally the genetic variants identified by GWAS only have small to modest

effects on the trait of interest in each individual, those variants contribute to the overall variation of that trait in the whole population [128]. Many studies have been done using GWAS to study important traits in livestock, especially quantitative traits. For example, Fortes et al. [129] tested the associations of SNP to 22 traits related to age at puberty, and captured some previously experimentally validated binding sites and identified new candidate genes and interactions. Xu et al. [22] used the BovineSNP50 assay to identify CNV associated with milk production traits in Holsteins by performing a conventional GWAS with SNP and characterizing the LD between SNP and CNV haplotypes to identify 34 CNV significantly associated with at least one milk trait.

One limitation for algorithms using WGS, however, is that CNV haplotype are unknown, making it difficult to do GWAS with CNVR directly. Because there are only limited studies about GWAS using CNVR with unknown CNV haplotype, alternative methods are incorporated. For example, by coding CNVR as a binary marker the extent of LD between CNVR and SNP can be determined, and then the impact of SNP having high LD with those CNVR can be analyzed [130, 131]. SNP associated with CNVR may then be used in GWAS and predictive modeling to study the impacts of CNVR on phenotypes.

In this study, two sets of bovine CNVR from a mapping population were used, LD with SNP were calculated, and SNP associated with those CNVR ($r^2 \geq 0.8$) were used to do GWAS. Results were compared to verify if CNVR was significantly associated with a trait of interest and could be used as a marker for selection in the future.

IV.2 Materials and Methods

IV.2.1 CNVR and SNP sets used in this study

Copy number variant regions were identified from the bam files of seven Nellore bulls and six Angus cows, which are the founders of the McGregor Genomics beef cattle population [85], as shown in II.2.1. Two sets of biallelic CNVR with odd copy numbers on BTA1 to BTA29 from II.4 and III.3.2 were summarized and used in this study: 1) CNVR set 1: the consensus CNVR set identified by CNV-seq [30] with 25 kb minimum detectable CNVR size, which were also detected by BreakDancer [47] and RAPTR-SV [52]; 2) CNVR set 2: the consensus CNVR set identified by three different control animals: A_ref, A03 and N05 using CNV-seq with 25 kb minimum detectable CNVR size.

The SNP were obtained from genotypes imputed to 770K array scale from the McGregor Genomics beef cattle population [85]. Three models: additive model, dominance model, and recessive model were used to code the SNP, as shown in Appendix B, Table B-6.

IV.2.2 Phenotype records used in this study

Phenotype data of 995 animals from McGregor Genomics beef cattle population [85] were used in this study, which include the ID, sex, birth date, weaning date, birth weight and weaning weight of each animal. Phenotypes data were cleaned and processed to have the following variables: ID, sex (three fixed levels: F, S and B), year and season of birth (random variable with various levels), weaning age (numeric covariate), birth weight (numeric response variable 1), and weaning weight (numeric response variable

2), as shown in Appendix C, Additional File C-19. Birth weight and weaning weight followed an approximate normal distribution (Appendix A, Figure A-9).

IV.2.2 CNVR coding and LD calculation of CNVR with SNP

In this study, we assume that the control animals have two copies in all copy number variant regions. Only biallelic CNVR with odd copy numbers on BTA1 to BTA29 were used (i.e. the sex chromosomes were omitted). In CNV-seq, the output $\log_2$ ratio is the log of the copy number ratio of the test animal to control animal [30], so the relative copy number (RCN) in the founders was calculated from $\mathrm{RCN} = 2^{(\log_2 \mathrm{ratio}+1)} - 2$, and rounded to the nearest integer. To be able to treat the biallelic CNVR as if they were SNP, they were then coded based on the following: AT for loss (-X), TT for normal (0) and CT for gain (X) [130, 131]. X is the copy number for insertions and deletions.

Linkage disequilibrium between CNVR and SNP in the range from 1 Mb upstream or downstream of that CNVR were then calculated using Plink [132, 133]. The threshold of r^2 was set to be ≥ 0.8. Finally, a list of SNP from each of the additive, dominant and recessive models having $r^2 \geq 0.8$ were selected for each of the two sets of CNVR for the subsequent analysis. These results were compared to those from using all SNP in GWAS for additive, dominance and recessive models.

IV.2.2 GWAS with the CNVR-SNP sets on birth and weaning weight

Proc Mixed in SAS was used to model the linear mixed regression to obtain residuals of birth and weaning weight of this population to be used in GWAS. For birth weight, sex was treated as a fixed factor, the year and season of birth and its interaction

with sex were treated as random factors. For weaning weight, gender and weaning age were treated as fixed factors, and the year and season of birth was treated as a random factor. Other interactions were ignored because they did not explain any of the variance.

For each CNVR-SNP set, GEMMA [134, 135] was then used for GWAS with phenotypes being the residuals of birth weight and residuals of weaning weight, respectively. The genomic relatedness matrix was calculated and association tests with Univariate Linear Mixed Models were used to perform Wald tests. The Wald p-values after Benjamini-Hochberg correction [136, 137] were then used to calculate $-\log_{10}(\mathrm{p}-\mathrm{value})$, which were plotted in Circos [138]. The proportion of variance in phenotype explained by a given SNP (PVE) is calculated following [139, 140].

IV.3 Results and discussion

IV.3.1 LD calculation between CNVR and SNP

The two sets of CNVR and their RCN are summarized in Appendix C, Additional File C-20, with CNVR set 1 having 113 CNVR and CNVR set 2 having 163 CNVR.

The six CNVR-SNP sets are shown in Table IV.1 and Appendix C, Additional File C-21. The number of CNVR associated with SNP and the number of those SNP for each model are also summarized in Table IV.1. The IDs of CNVR associated with each SNP model and their location in the genome are listed in Appendix C, Additional File C22 for each of the CNVR-SNP sets. For CNVR set 1, 83.19% of the CNVR were tagged by SNP; for CNVR set 2, 91.41% of the CNVR were tagged by SNP.

Table IV.1 The SNP sets used in this study. Only CNVR and SNP on BTA1 – BTA29 were included.

CNVR-SNP set	SNP association with CNVR set	SNP model	Number of total CNVR associated with SNP	Number of total SNP in association
1	CNVR set 1	Additive	94	42595
2	CNVR set 1	Dominant	88	16347
3	CNVR set 1	Recessive	66	11476
4	CNVR set 2	Additive	149	108428
5	CNVR set 2	Dominant	121	17690
6	CNVR set 2	Recessive	93	13372

IV.3.2 GWAS for birth weight and weaning weight

Genome-wide association studies were done for birth and weaning weight for each of the CNVR-SNP sets and these were compared to GWAS using all SNP from additive, dominance and recessive models. The *P*-values of the significant SNP after Benjamini-Hochberg correction are summarized in Appendix C, Additional File C-23 and C-24 for birth weight and weaning weight, respectively. Plots of the $-\log_{10} P-$ values after Benjamini-Hochberg correction are shown in Figure IV.1 and Figure IV.2, and circus plots of the $-\log_{10} P-$ values before Benjamini-Hochberg correction are shown in Appendix A, Figure A-10. After Benjamini-Hochberg correction, most SNP marking the CNVR associated with the traits were not significant and so they were not plotted. For the SNP remaining in the tracks, some had very small *P*-values, which indicates there is evidence in those regions for differences in allele frequency affecting the respective phenotype. The PVEs are summarized in Appendix C, Additional File C-23 and C-24. Some of the SNP marking CNVR were significant for different SNP models, and some were the same as in the GWAS track having all SNP in that model. Some GWAS peaks, however, appeared in the tracks for the CNVR-SNP sets, but did not appear in the tracks for all SNP for the respective models. Different models had some of the same significant SNP as in GWAS. For example, the peaks on BTA2 and BTA21 for birth weight, and BTA12 for weaning weight, indicating these regions might affect the respective phenotype.

For birth weight with CNVR-SNP set 1 (Figure IV.1 a), peaks appeared on BTA10 and BTA12; for birth weight with CNVR-SNP set 4 (Figure IV.1 a), peaks

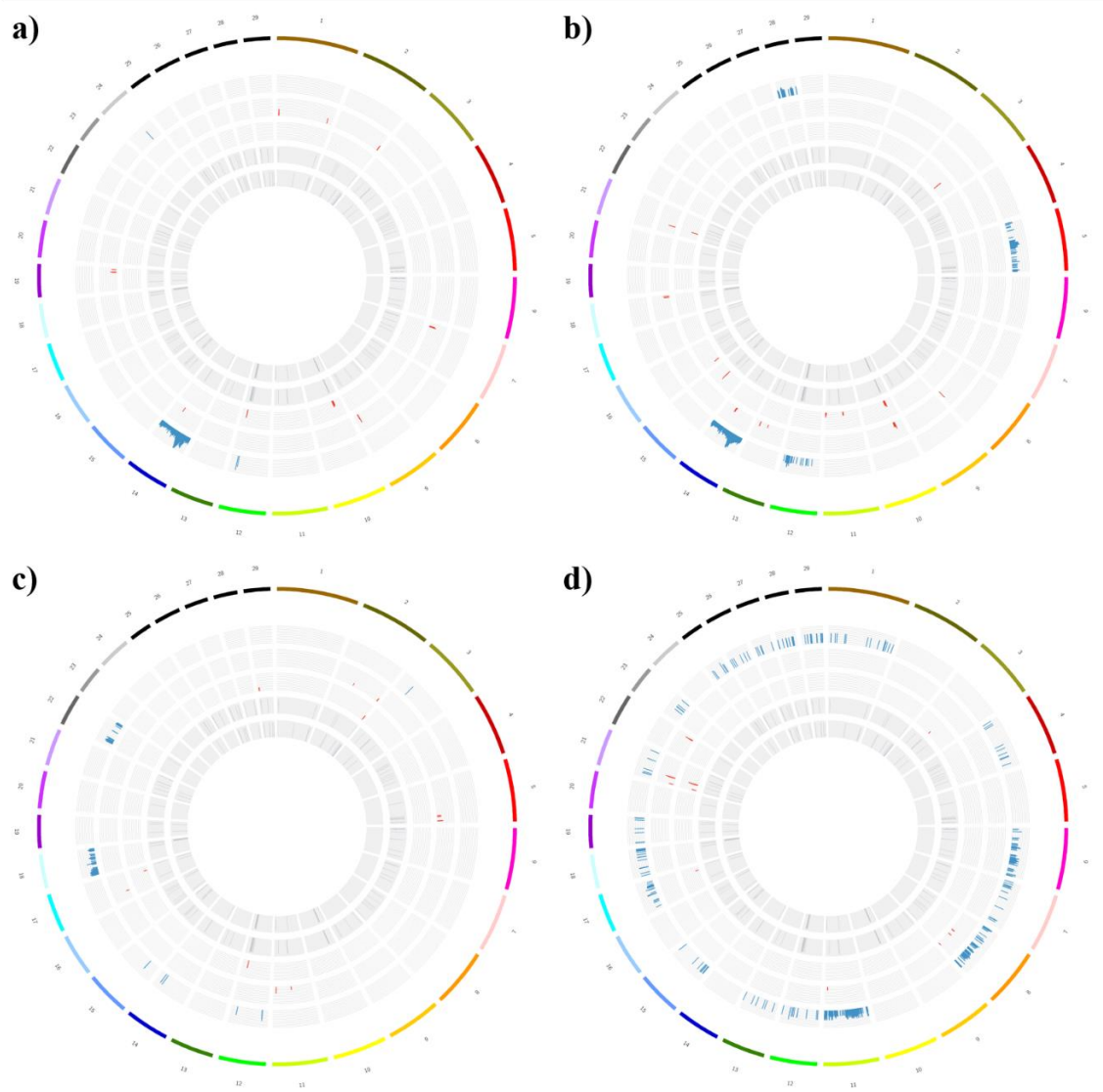


Figure IV.1 Circos plots of $-\log_{10}$ P for CNV tagging SNP. Circos plots of $-\log_{10}$ P for CNV tagging SNP associating with birth weight and weaning weight for the 6 CNVR-SNP models on BTA1-BTA29 after Benjamini-Hochberg correction are plotted. For each circus plot, the tracks from inside to outside are: CNVR set 1, CNVR set 2, SNP associating with CNVR set 1 and CNVR set 2 for a-b) additive SNP model (CNVR-SNP set 1 and 4), c-d) dominance SNP model (CNVR-SNP set 2 and 5), and e-f) recessive SNP model (CNVR-SNP set 3 and 6), and all SNP in that SNP model, respectively. Figure a), c) and e) are for birth weight and Figure b), d) and f) are for weaning weight. For CNV tracks, red indicate insertions and blue indicate deletions.

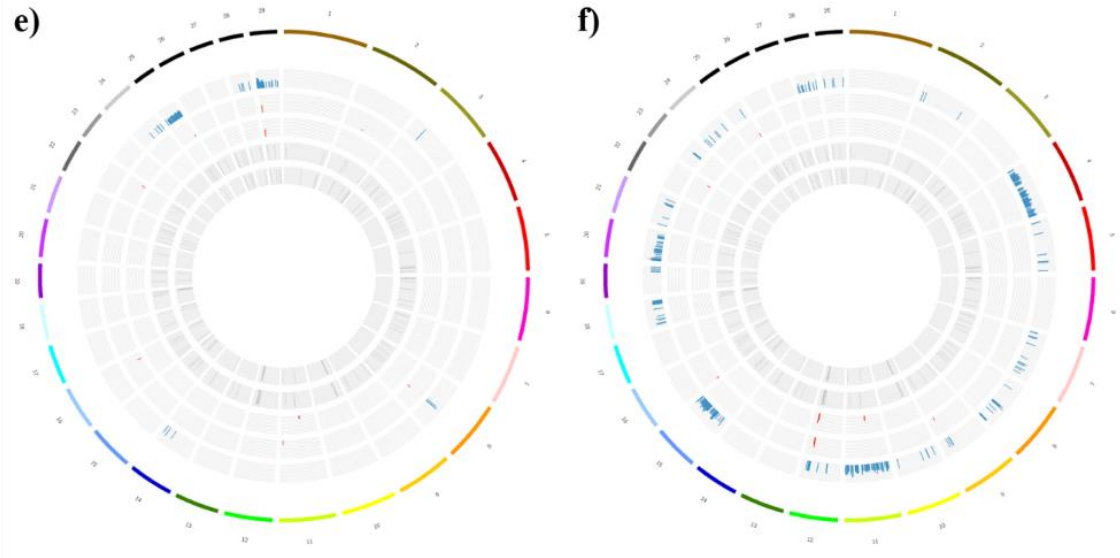


Figure IV.1 Continued.

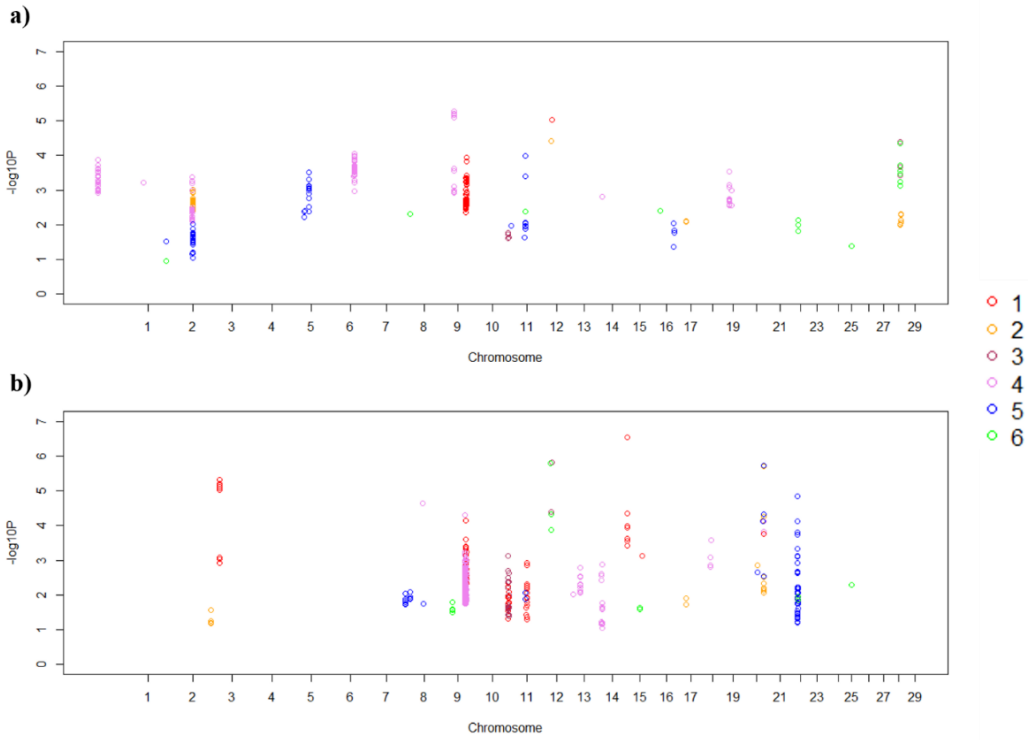


Figure IV.2 Manhattan plots of $-\log_{10}$ P for SNP. Manhattan plots of $-\log_{10}$ P for SNP associated with a) birth weight and b) weaning weight for the 6 CNVR-SNP models on BTA1-BTA29. Each color represents a different CNVR-SNP set as shown in the legend.

appeared on BTA1, BTA2, BTA7, BTA9, BTA14, and BTA19. The peaks on BTA12 with set 1 and BTA14 with set 4 also appeared at similar loci in the GWAS track for all SNP in the additive model. For weaning weight with CNVR-SNP set 1 (Figure IV.1 b), peaks appeared on BTA3, BTA10, BTA11, BTA15, BTA16 and BTA21; for weaning weight with CNVR-SNP set 4 (Figure IV.1 b), peaks appeared on BTA8, BTA10, BTA13, BTA14, BTA18 and BTA21. The peak on BTA14 also appeared at similar loci in the GWAS track of all SNP for the additive model. The peaks on BTA10 and BTA21 appeared for both CNVR-SNP sets 1 and 4, but did not show up in the track of GWAS for all SNP for the additive model, indicating that this region may have a high possibility of a CNV affecting weaning weight.

For birth weight with CNVR-SNP set 2 (Figure IV.1 c), peaks appeared on BTA2, BTA12, BTA17 and BTA29. For birth weight with CNVR-SNP set 5 (Figure IV.1 c), peaks appeared on BTA2, BTA5, BTA11 and BTA17. The peak on BTA17 also appeared at similar loci in the GWAS track having all SNP under the dominance model. One peak on BTA2 appeared for both CNVR-SNP set 2 and 5, but did not show up in the track of GWAS for all SNP under the dominance model. For weaning weight with CNVR-SNP set 2 (Figure IV.1 d), there were peaks on BTA3, BTA17 and BTA21. For weaning weight with CNVR-SNP set 4 (Figure IV.1 d), there were peaks on BTA8, BTA11, BTA21 and BTA22. The peaks on BTA8, BTA11 and BTA21 also appeared at similar loci in the GWAS track of all SNP for the dominance model. In addition, one peak on BTA21 appeared for both CNVR-SNP set 2 and 5 and with all SNP, indicating that this region may have a high probability of affecting weaning weight.

For birth weight with CNVR-SNP set 3 (Figure IV.1 e), peaks only appeared on BTA11 and BTA29. For birth weight with CNVR-SNP set 6 (Figure IV.1 e), peaks appeared on BTA2, BTA8, BTA11, BTA16, BTA22, BTA25 and BTA29. The peak on BTA29 appeared for both CNVR-SNP set 3 and 6, and appeared at similar loci in the GWAS track of all SNP under the recessive model. For weaning weight with CNVR-SNP set 3 (Figure IV.1 f), peaks only appeared on BTA11 and BTA12. For weaning weight with CNVR-SNP set 6 (Figure IV.1 f), peaks appeared on BTA9, BTA12, BTA16, BTA22 and BTA25. The peaks on BTA12 appeared for both CNVR-SNP set 3 and 6, and appeared at similar loci in the GWAS track of all SNP under the recessive model.

IV.3.3 RefSeq genes overlapping with significant SNP and CNV, and comparison to other studies

The RefSeq genes overlapping with CNVR-tagged SNP are summarized in Table IV.2 and Appendix C, Additional File C-25. The genes overlapping with CNVR-SNP set 1 to 6 range from 3 to 12 for birth weight and 3 to 31 for weaning weight. There were 30 and 56 genes overlapping with CNVR-tagged SNP for birth weight and weaning weight, respectively, and 4 genes were in common: T cell receptor delta chain variable region BVd1.15 (BVD1.15 or BVD1.18), ubiquitin specific peptidase 20, LIM domains containing 1, and uncharacterized LOC100336282. Among them, ubiquitin specific peptidase 20 was found to be involved in bovine respiratory disease by a SNP association study from Neupane et al. [141], and LIM domains containing 1 was found

Table IV.2 Number of RefSeq genes overlapping with significant SNP in the 6 CNVR-SNP sets.

Phenotype	CNVR-SNP set	Number of RefSeq genes overlapped
Birth weight	1	3
	2	9
	3	1
	4	11
	5	12
	6	3
Weaning weight	1	15
	2	3
	3	7
	4	31
	5	8
	6	3

to be associated with head and neck squamous cell carcinoma in human according to Ghosh et al. [142].

No genes were found by all 3 GWAS models for birth weight or weaning weight. However, many genes were found by 2 of the 3 GWAS models. Taking both birth weight and weaning weight into account, BVD1.18 was found by most of the GWAS models with CNVR-SNP set 1 for birth weight and with CNVR-SNP sets 1 and 4 for weaning weight. The common genes discovered overlapping with CNVR-tagging SNP in two of the CNVR-SNP sets for birth weight and weaning weight are summarized in Table IV.3. Among them, CHORDC1 on BTA29 was found by Anton et al. [143] to be associated with breeding value of beef. The SNP identified by Anton et al. is compared to SNP on BTA29 close to CHORDC1 identified in our study which associate with birth weight and weaning weight in Table IV.4.

Some genes discovered directly overlap with CNVR, including secreted and transmembrane protein 1A, WC1.3 molecule, and protein tyrosine phosphatase, receptor type T (PTPRT), as shown in Table IV.5. None of the genes indicated above were previously reported to affect birth weight or weaning weight. However, some studies showed that PTPRT is included in a microdeletion in human, and this CNVR may be causative for neurodevelopmental disorders [144, 145].

IV.4 Conclusions

SNP tagging CNVR can be used as genetic markers in GWAS to identify significant SNP and CNVR, and important genes associated with them. Focusing just on the SNP tagging specific CNVR reduces the multiple testing problem. Different SNP

Table IV.3 The common genes discovered overlapping with CNVR-tagging SNP in two of the CNVR-SNP sets.

Pheno-type	Chromo-some	Gene	Identification in other SNP association studies
Birth weight	2	Peptidyl arginine deiminase 2 (PADI2)	Rheumatoid arthritis in human and angiogenesis-regulation in mice [146, 147]
	2	F-box protein 42 (FBXO42)	-
	2	EPH receptor A2 (EPHA2)	Cataract in human [148, 149]
	2	Family with sequence similarity 131 member C (FAM131C)	-
	2	Chloride channel Ka (CLCNKA)	Heart failure and glomerular filtration in human [150]
	2	Heat shock protein family B (small) member 7 (HSPB7)	Idiopathic dilated cardiomyopathy in human [151]
	29	Cysteine and histidine rich domain containing 1 (CHORDC1)	**Breeding value of beef** [143] and bovine respiratory disease [141]
Weaning weight	10	T cell receptor alpha variable 14/delta variable 4 (505306)	-
	10	T cell receptor delta chain variable region BVd1.15 (BVD1.18)	-
	21	Neurotrophic receptor tyrosine kinase 3 (NTRK3)	Several neural disorders in human [152-157]
	21	Aggrecan (ACAN)	-

Table IV.4 Comparison of SNP on BTA29 close to CHORDC1 identified by Anton et al. and our study.

CNVR-SNP set	chromosome	Position	MAF	-log10P	PVE
3	29	5057550	0.037	4.399647	0.016824
	29	5057695	0.037	4.399647	0.016824
	29	5058080	0.037	4.399647	0.016824
	29	4483843	0.185	3.671313	0.01369
	29	4477357	0.104	3.469031	0.012824
	29	4465341	0.109	3.424037	0.012631
6	29	5057550	0.037	4.346115	0.016593
	29	5057695	0.037	4.346115	0.016593
	29	5058080	0.037	4.346115	0.016593
	29	4483843	0.185	3.716973	0.013885
	29	4477357	0.104	3.5575	0.013202
	29	4465341	0.109	3.471724	0.012835
	29	4927893	0.077	3.242844	0.011858
	29	4917583	0.079	3.12765	0.011368
Anton et al.	29	3901625	0.354	14.5	-

Table IV.5 RefSeq genes overlapping with significant SNP that have direct overlap with CNVR. BW indicate birth weight and WW indicate weaning weight.

CNVR-SNP set	Phenotype	Chromosome	Start	End	Type	Gene
1	WW	10	23543751	23571250	Deletion	LOC100335575
2	WW	21	20158751	20188750	Deletion	Myeloid-associated differentiation marker-like (LOC618633)
4	WW	10	22173751	22323750	Insertion	LOC100336282
4	WW	10	23543751	23571250	Deletion	LOC100335575
4	WW	13	71903751	71933750	Deletion	protein tyrosine phosphatase, receptor type T (PTPRT)
		18	61703751	61731250	Deletion	cationic amino acid transporter 3
4	BW	19	51036251	51068750	Deletion	Secreted and transmembrane protein 1A (SECTM1A)
5	BW	5	103163751	103223750	Insertion	WC1.3 molecule (WC1.3)

models yielded different results; however, different SNP models had some of the same significant peaks in GWAS, indicating these regions might affect the respective phenotype. Some peaks appeared at similar loci in the GWAS track of all SNP for the respective SNP model, indicating that the findings using CNVR yielded similar results to studies solely with SNP. Some peaks appeared for one or multiple CNVR-SNP sets but not with SNP alone, indicating that CNV in these regions may affect birth or weaning weights.

CHAPTER V

PREDICTION OF EFFECTS OF CNVR ON BIRTH AND WEANING WEIGHT IN A BEEF CATTLE MAPPING POPULATION

V.1 Introduction

One approach to analyze the association of genetic variants with phenotypes is GWAS, as shown in Chapter IV. However, although GWAS can detect the genetic variants which are strongly associated with interesting traits, these variants typically only explain a small portion of the genetic variance and heritability, and they hence have low predictive power [158, 159]. An alternative approach is to use the genetic variants across the genome to predict interesting traits. Phenotypic prediction is typically performed with statistical models, and this is the approach we take here.

One frequently used model is the Bayesian sparse linear mixed model (BSLMM) [160], which is a hybrid of a linear mixed model (LMM) in which every genetic variant affects the phenotype and a sparse regression model in which only a small proportion of all variants affect the phenotype. It combines the advantages of both models, yields good performance across a wide range of genetic architectures, and is valid and stable for phenotype prediction [160]. Because we do not know how large a proportion of all variants affects the phenotype, this model is well suited for our data and was chosen. Since we aimed to predict both birth weight and weaning weight as phenotypic outcomes, we also considered a multivariate linear regression (MLR) model (a regression model in which the response is multivariate) [161]; MLR models are often

used in QTL mapping with pedigree data [162, 163]. The advantage of this model is that it is able to find genetic loci that influence multiple traits jointly [164]. This is the only model in our study that detects CNVR which influence both traits simultaneously. The regression tree (RT) model and random forest (RF) model are nonparametric machine-learning methods, with underlying theories quite different from the linear approaches described above. Therefore, they might be able to capture unique relationships between genetic variants and traits [165, 166]. The RT model is fit by recursively partitioning the data space and fitting simple prediction models within each partition, the results of which can be represented by a decision tree [166]. RFs are ensembles of classification and regression trees that can predict the outcome based on a large number of predictors like SNP [167, 168]. An example of RF applied in SNP analysis is shown in [169].

In 2015, Moser et al compared the performance of BSLMM, Hierarchical Bayesian Mixture Model (BayesR), LMM and a single-SNP analysis, and concluded that the Bayesian models had the highest true positive rate and prediction accuracy [159]. In 2017, Zeng et al. compared BSLMM model to Lasso, Lasso and elastic net and LMM, and concluded that BSLMM performed best across different scenarios [170]. In 2000, Comings et al. used a multivariate regression model to simultaneously analyze the effect of 20 genes on a range of phenotypes for attention deficit hyperactivity disorder in 336 unrelated Caucasian subjects [171]. In 2009, based on multivariate regression with 53 clinical traits related to severe asthma and 34 SNP from 543 asthma patients, Kim et al. developed and compared the performance of several lasso regression models [164]. In 2010, Peng et al. also developed a regularized multivariate regression model for

identifying master predictors for breast cancer analysis [172]. The RT and RF models are widely used in genetic analysis including large SNP data sets from GWAS [173-175]. In 2009, García-Magariños et al. recommended tree based models for large-scale genetic data where there are unknown interactions among true risk-associated SNP with marginal effects, and where a significant number of SNP due to noise are present [175]. In 2010, Goldstein et al. performed RF on a case-control dataset with 300,000 SNP genotypes across the genome, and concluded that RF is computationally feasible for GWA data, and the results made biologic sense [174]. However, although RF is widely used with large SNP data, Winham et al. argued that as dimensionality increased, RF's detection ability declined more rapidly for interacting SNP than for non-interacting SNP [169].

In this study, four different models, BSLMM, MLR, RT and RF were used to predict birth weight and weaning weight in a beef cattle mapping population. Two sets of bovine CNVR which were tagged by SNP were used as training data. Three different SNP models: additive, dominant and recessive, which were coded as in [176] and Appendix B, Table B-6, were tested. Predictive accuracies were assessed, and model performance for each set of CNVR were compared. The study is novel because few phenotype prediction studies have been done with beef cattle previously, and a novel approach to phenotype prediction using CNVR and associated SNP was proposed, which makes it possible to predict phenotypes in a large population based on CNVR detected from WGS data in a much smaller population. In addition, four different models with distinct underlying theories were performed and compared to provide guidance in

choosing predictive models for livestock. This is broadly applicable to the field of breeding and selection in livestock, and, by extension, even in human medicine.

V.2 Methods

V.2.1 Source of data

The six CNVR-SNP sets from IV.3.1 were used in this study. SNP in additive, dominant and recessive models were further coded for MLR, RF and RT models as in Appendix B, Table B-7. Phenotype data is the same as IV.2.2, except that it was processed to have the following variables: ID, sex (three fixed levels: F, S and B), birth season (4 fixed levels: spring, summer, autumn and winter), weaning age (numeric covariate), birth weight (numeric response variable 1), and weaning weight (numeric response variable 2), as shown in Appendix C, Additional File C-19. The response variables and weaning age were standardized in the following studies. The standardized response variables and their residuals from a MLR model with sex, birth season and weaning age were approximately normally distributed, as shown in Appendix A, Figure A-11.

V.2.2 Predictive modeling

The mapping population (n = 995 with non-missing phenotypes) was randomly split to training (n = 500) and testing (n = 495) data sets. Each of the 6 CNVR-SNP sets was tested for their performance for prediction of birth weight and weaning weight. In order to test the prediction performance of CNVR to the 2 phenotypes, SNP lists from each of the 6 CNVR-SNP sets were split based on which CNVR they were associated

with and were used for predictive modeling. Therefore, the prediction effect of SNP associated with a specific CNVR represents the prediction effect of that CNVR.

Four predictive models were used: (1) BSLMM with a ridge regression / genomic best linear unbiased prediction (GBLUP) with standard non- Markov chain Monte Carlo (MCMC) method in GEMMA [160], (2) MLR, (3) RT and (4) RF. For each model, in the training data set, a 10-fold cross-validation (CV) was done to rank the CNVR based on their prediction accuracy (quantified by mean squared error (MSE)). This was followed by another round of 10-fold CV to find the number of ranked CNVR that yielded the highest prediction accuracy. The models were then assessed by prediction on the testing data set. Adjusted R^2 and the Bayesian Information Criterion (BIC) [177] were also calculated for model comparison purposes. As one additional guide for model selection, we calculated Pearson correlation between predicted and original values of birth and weaning weights. The predicted birth and weaning weights were obtained by transferring back the residuals or standardized values to the original scale.

The bootstrap [178] was performed to assess the variability of model performance estimates. Standard nonparametric bootstraps were performed for models (2), (3) and (4). However, for the BSLMM model, due to limitations of the GEMMA software, the MSE in each bootstrap loop was obtained by using the first 2/3 of the randomized bootstrap data as training data to perform prediction on the last 1/3 of the randomized bootstrap data. Finally, the performance of these models were compared. In addition, MSEs of training and prediction steps from a MLR model with sex, birth season and weaning age only, which were referred to as baseline training MSEs and

baseline prediction MSEs, were used to assess the performance of the four models with SNP effects in the model. A flow chart of this study is shown in Figure V.1.

V.2.2.1 The BSLMM model

We used the Plink software [132, 133] to prepare the necessary binary files containing SNP information. Phenotype information (birth and weaning weight) were added to the corresponding fam files one at a time, with phenotype values needing prediction masked. Those files were then used to fit the BSLMM model and perform prediction in GEMMA. The ridge regression/GBLUP with standard non-MCMC method were used because of its fast computation time. Preliminary results showed that the MSEs obtained from this option were very close to the linear BSLMM using MCMC for our data (data not shown). Since the BSLMM model in GEMMA does not accept covariates, residuals from a MLR model with sex, birth season and weaning age only were used as phenotype values.

The BSLMM model in GEMMA is as follows [160]:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}$$

$$\beta_i \sim \pi \mathrm{N}(0, \sigma_a^2\tau^{-1}) + (1-\pi)\delta_0, \qquad \mathbf{u} \sim \mathrm{MVN}_n(0, \sigma_b^2\tau^{-1}\mathbf{K}), \qquad \boldsymbol{\epsilon} \sim \mathrm{MVN}_n(0, \tau^{-1}\mathbf{I}_n)$$

Where $\boldsymbol{y}$ is a vector of residuals of standardized birth or weaning weight, $\mathbf{I}_n$ is an n-vector of 1s, μ is a scalar representing the phenotype mean, $\mathbf{X}$ is an n × p matrix of genotypes measured on n individuals at p genetic markers, $\boldsymbol{\beta}$ is the corresponding p-vector of the genetic marker effects, with β_i for the ith column; $\mathbf{u}$ is the term of random effects for phenotype means; $\boldsymbol{\epsilon}$ is the term of random error; and MVN_n denotes the n-dimensional multivariate normal distribution.

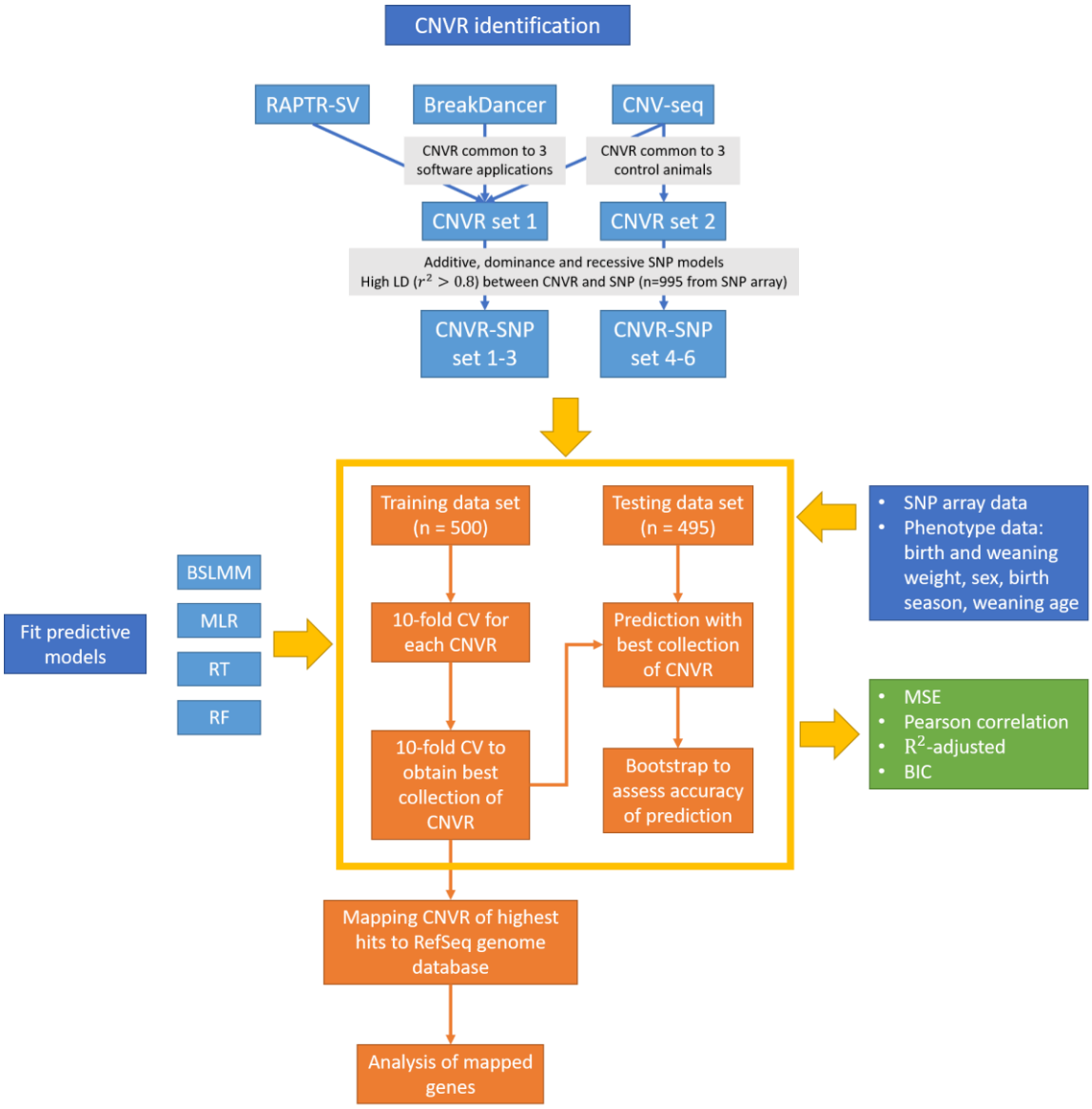


Figure V.1 Flow chart of the study.

The other three models were performed by custom scripts using plink and R [179].

V.2.2.2 The MLR model

The multivariate (vector response) linear regression model is as followings:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

Where $\mathbf{Y}$ is a $n \times 2$ vector which contains the responses (standardized birth and weaning weights). $\mathbf{X}$ contains the covariates for gender (categorical, with levels "F", "S" and "B"), birth season (categorical, with levels "spring", "summer", "autumn" and "winter"), and standardized weaning age; see Appendix C, Additional File C-26 for details of how the covariates were coded.

When fitting this model to the training data, SNP having exactly the same pattern and SNP with only one level were deleted. In cases where multicollinearity was an issue (design matrix computationally singular), a generalized linear model with lasso penalty [180] was used to shrink the model coefficients. During this step, sex, birth season and weaning age always remained in the model. As a result, only a very small subset of SNPs was selected by the MLR model. Two sets of coefficients were then selected: (1) coefficients based on the λ value that gives minimum mean CV error (lambda.min) and (2) coefficients based on the λ value from the most regularized model such that error is within one standard error of the minimum (lambda.1se). Two sets of SNP associated with CNVR were then selected based on the two sets of coefficients, and these were added to the design matrix. After this step, if multicollinearity was still an issue, the corresponding model's MSE was recorded as 'NA' (missing). Also, throughout the above

SNP filtering steps, if no SNP were left after filtering, further steps were not performed for that SNP set, and the model's MSE was recorded as 'NA'. For validation groups in the training data, the same SNP were selected in accordance with SNP selected in the training groups. Thus, validation group models have exactly the same covariates as training groups (same columns in the design matrix). Finally, if more than 5 MSEs were missing values for any 10-fold CV, a missing value was assigned to its overall MSE. In the end, only the results from lambda.min were used in the following steps because it produced smaller MSEs and fewer missing values for the MSEs.

V.2.2.3 The RT and RF models

Finally, we employed two decision tree-based models: (1) a basic RT (using the "tree" package in R [181] with default settings) and (2) a RF model (using the "randomForest" package in R [182] with default settings). The default parameters for the RF model were chosen because preliminary studies with 10 randomly selected CNVR from CNVR-SNP set 1 showed that increasing the number of trees to grow ('ntree') beyond the default value (50) did not substantially change model accuracy (Appendix A, Figure A-12). For the number of variables to choose per node (mtry), the default for classification trees is $\sqrt{\text{number of SNP}}$ and the default for regression trees is the number of SNP / 3. Although the former performed slightly better than the latter, we still chose to use the latter in our study since our response variables were numeric. Since we want to keep the effects of sex, birth season and weaning age in our models, residuals from a MLR model with sex, birth season and weaning age only were used as phenotype values for birth and weaning weight for both of the models.

V.2.3 Identification of genes overlapping with best collections of CNVR

Custom perl scripts were used to identify the genes overlapping with best collections of CNVR for each of the models. The RefSeq *Bos taurus* (assembly Bos taurus_UMD_3.1) genome database was used. InteractiVenn [98] was used to obtain the common genes identified by the four models.

V.3 Results and discussion

V.3.1 Building and tuning models with training data and CV

The MSE obtained for each CNVR of the 6 CNVR-SNP sets from 10-fold CV with the four models is summarized in Figure V.2. It can be seen that BSLMM model yields relatively smaller MSEs compared to other models. Almost all of its MSEs were lower than baseline training MSEs (0.8864 for birth weight and 0.6004 for weaning weight). The other three models have a relatively large proportion of MSE values above the baseline training MSEs.

The MSEs for varying numbers of ranked CNVR in each CNVR-SNP set for each model were then obtained and summarized for each of the models. The summarization for the BSLMM model is shown in Figure V.3 (a). For both birth and weaning weight, the MSEs first decrease and then increase as a function of the number of CNVR used as predictors. All MSEs obtained were less than the baseline training MSEs. The number of ranked CNVR corresponding to the minimum MSE for BSLMM model, and their model fit statistics are shown in Table V.1. For birth weight, the lowest MSE obtained from the training data set was 0.74 from CNVR-SNP set 4; 43 out of 144 ranked CNVR were used to obtain this MSE. For weaning weight, the lowest MSE

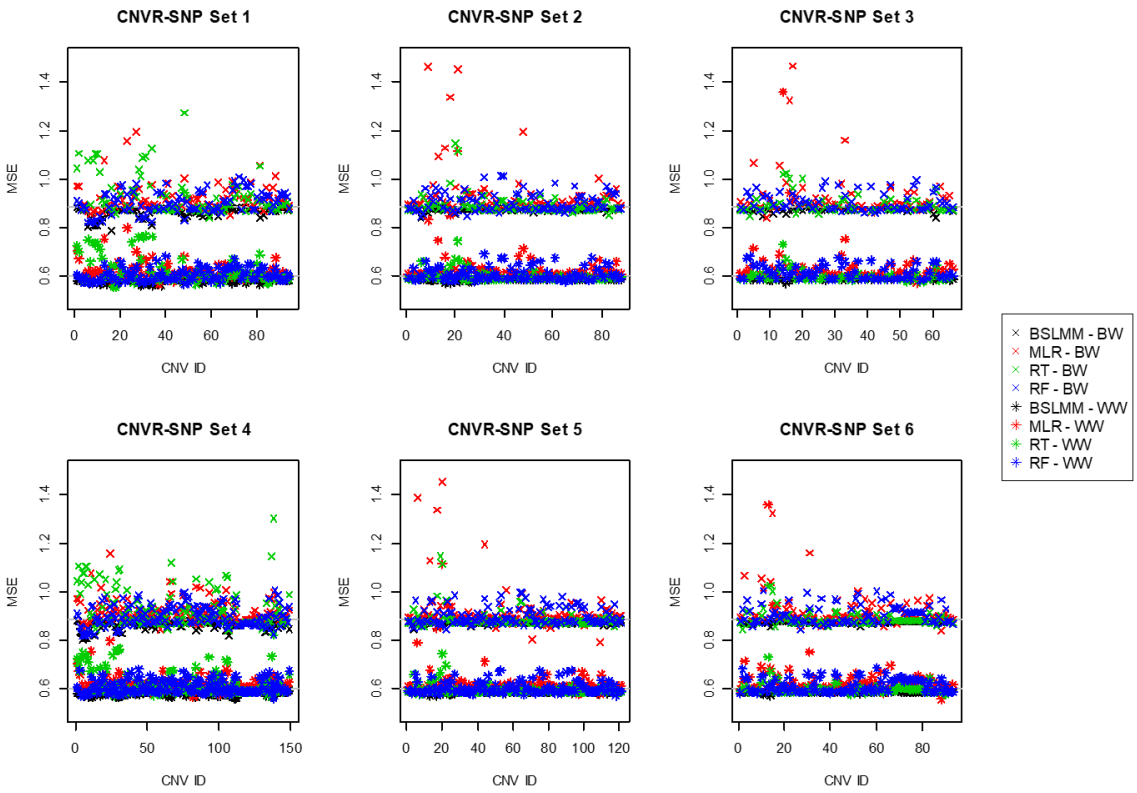


Figure V.2 MSE of each CNVR from CV with the four models. The MSE of each CNVR from CV with a) BSLMM, b) MLR, c) RT and d) RF models are summarized for each CNVR-SNP set. The horizontal grey lines are baseline MSEs for predicting birth weight and weaning weight, but may not be visible because there are too many MSEs close to baseline MSEs.

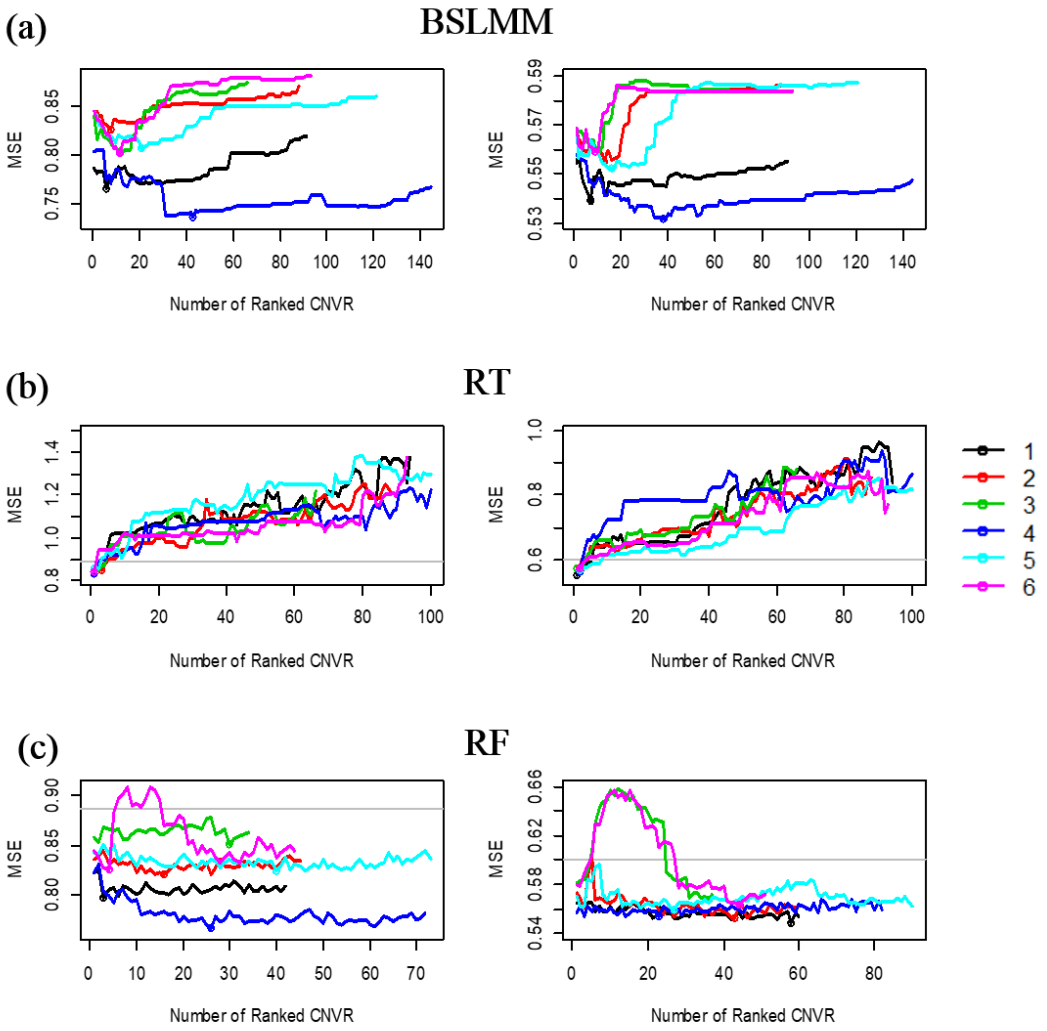


Figure V.3 MSEs of varying numbers of ranked CNVR from CV. The MSEs of accumulating lists of ranked CNVR from CV for a) BSLMM b) RT and c) RF models are summarized for each of the CNVR-SNP set. 1-6 indicates CNVR-SNP set 1-6. The left side is MSEs for prediction of birth weight, and the right side is MSEs for prediction of weaning weight. The horizontal grey lines in (b) and (c) indicate baseline MSEs for birth and weaning weights, respectively. All MSEs in (a) were below baseline MSEs. For RT, only the first 100 (or the total if total were less than 100) are displayed.

Table V.1 Number of ranked CNVR to obtain minimum MSE for BSLMM model.

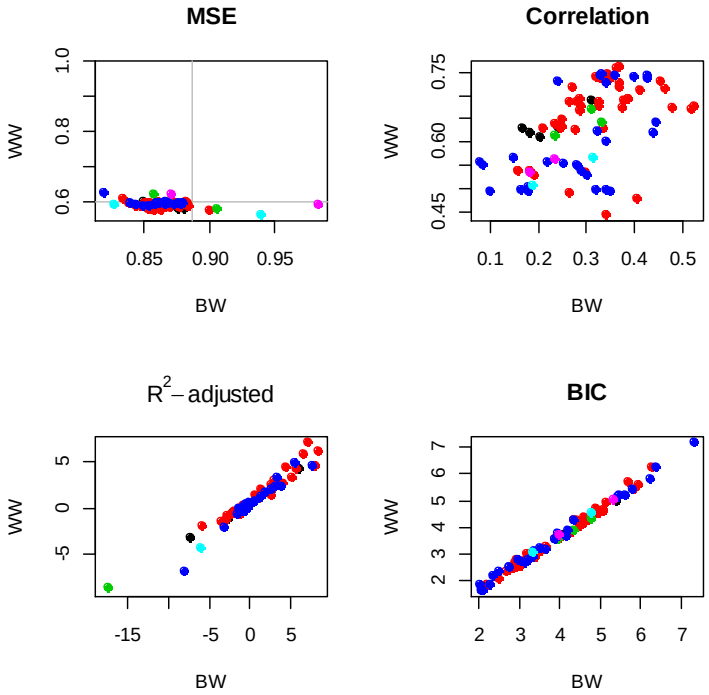
Phenotype	CNVR-SNP set	Total number of valid CNVR	Best number of ranked CNVR	Minimum MSE	Correlation	Adjusted R^2	BIC
Birth weight	1	91	6	0.76	0.45	1.03	124.86
	2	88	8	0.83	0.35	1.03	139.90
	3	66	12	0.80	0.39	1.02	213.68
	4	144	43	0.74	0.48	1.00	3557.32
	5	121	21	0.81	0.38	1.02	236.48
	6	93	12	0.80	0.40	1.03	148.14
Weaning weight	1	91	7	0.54	0.68	1.02	189.04
	2	88	13	0.56	0.67	1.04	109.00
	3	66	10	0.56	0.67	1.03	150.50
	4	144	38	0.53	0.69	1.00	3523.72
	5	121	16	0.55	0.68	1.04	106.18
	6	93	9	0.56	0.67	1.03	134.40

The best number of ranked CNVR to obtain minimum MSE from CV for BSLMM model is summarized for each of the CNVR-SNP set. Their corresponding correlation, R^2 – adjusted and BIC are summarized as well. The best CNVR-SNP sets are marked in bold.

obtained from the training data set was 0.53, which was again from CNVR-SNP set 4; 38 out of 144 ranked CNVR were used to obtain this MSE. Note that the MSEs for weaning weight prediction tended to be smaller than those for birth weight prediction. For the MLR model, there was no systematic association between the number of CNVR used and prediction accuracy (figure not shown). In most cases, their MSEs were larger than the baseline training MSEs. We obtained lists of varying numbers of ranked CNVR for each of the CNVR-SNP sets, which included the CNVR lists having at least one of the birth or weaning weight MSEs minimized (none of the lists have MSEs minimized for both birth and weaning weights simultaneously), or having both birth and weaning weight MSEs smaller than the baseline training MSEs, as summarized in Figure V.4 (b) and Appendix B, Table B-8. Their MSE, correlation, $R^2 - \text{adjusted}$ and BIC values are summarized in Figure V.4 (a) as well. The number of lists for different CNVR-SNP sets varies a lot. In most cases, only a few CNVR were selected for use in the MLR model, indicating that the MLR model may not be a very good fit for this population.

For the RT model, the MSEs for varying numbers of ranked CNVR in each CNVR-SNP set were summarized in Figure V.3 (b), and the number of ranked CNVR to obtain minimum MSE along with their model fit statistics were summarized in Table V.2. For birth weight, the lowest MSE obtained from the training data set was 0.84 from CNVR-SNP set 4; 1 out of 149 ranked CNVR were used to obtain this MSE. For weaning weight, the lowest MSE obtained from the training data set was 0.55, which was again from CNVR-SNP set 1; 1 out of 94 ranked CNVR were used to obtain this MSE. For this model, the MSEs tend to increase with increasing numbers of CNVR, and

(a)



(b)

CNVR-SNP set	Analyzed CNVR	CNVR collection to minimize MSE for BW	CNVR collection to minimize MSE for WW	Number of CNVR collections having MSEs < baseline MSEs of both BW and WW
1	33*46 of 90*90	2*1	1*2	5
2	13*43 of 81*81	7*18	1*2	40
3	19*25 of 62*62	16*10	1*2	1
4	29*26 of 141*141	23*5	10*14	27
5	39*70 of 113*113	1*1	4*1	1
6	26*41 of 87*87	1*1	2*4	0

Figure V.4 Best CNVR to obtain minimum MSE for MLR model. The best collections of CNVR to obtain minimum MSEs for birth and/or weights, and MSEs smaller than baseline MSE for birth and weaning weights simultaneously for MLR model are summarized. Their MSE, correlation, adjusted R^2 and BIC are summarized as well. In b, the numbers before and after '*' are best collection of ranked CNVR based on birth weight and best collection of ranked CNVR based on weaning weight, respectively. Note that due to computation time, for CNVR-SNP set 4, only CNVR with MSEs less than baseline MSEs were analyzed (0.8864 for birth weight and 0.6004 for weaning weight); for other CNVR-SNP sets, CNVR with MSE less than 0.89 for birth weight and 0.61 for weaning weight were analyzed. BW indicate birth weight and WW indicate weaning weight.

Table V.2 Number of ranked CNVR to obtain minimum MSE for RT model.

Phenotype	CNVR-SNP set	Total number of valid CNVR	Best number of ranked CNVR	Minimum MSE	Correlation	Adjusted R^2	BIC
Birth weight	1	94	1	0.84	0.33	2.74	7.02
	2	88	3	0.85	0.34	1.40	14.45
	3	66	1	0.84	0.33	1.09	48.80
	4	149	1	0.84	0.33	1.08	51.38
	5	121	1	0.86	0.32	1.06	68.84
	6	93	1	0.84	0.32	0.00	1.06
Weaning weight	1	94	1	0.55	0.67	1.50	12.15
	2	88	1	0.57	0.65	-0.03	1.08
	3	66	1	0.57	0.66	-0.33	1.94
	4	149	2	0.56	0.66	1.35	15.62
	5	121	2	0.57	0.66	-0.04	1.15
	6	93	2	0.57	0.66	1.17	27.53

The best number of ranked CNVR to obtain minimum MSE from CV for RT model is summarized for each of the CNVR-SNP set. Their corresponding correlation, $R^2 - \text{adjusted}$ and BIC are summarized as well. The best CNVR-SNP sets are marked in bold.

the majority of these MSEs were larger than the baseline training MSEs, for both birth and weaning weight. In addition, although all minimum MSEs from the 6 CNVR-SNP sets were smaller than the baseline training MSEs, only 1 to 3 CNVR were utilized, indicating the RT model may not be a very good fit to this population.

Finally, for the RF model, the MSEs for varying numbers of ranked CNVR in each CNVR-SNP set were summarized in Figure V.3 (c), and the number of ranked CNVR to obtain minimum MSE along with their model fit statistics were summarized in Table V.3. For birth weight, the lowest MSE obtained from the training data set was 0.77 from CNVR-SNP set 4; 26 out of 72 ranked CNVR were used to obtain this MSE. For weaning weight, the lowest MSE obtained from the training data set was 0.55, which was again from CNVR-SNP set 2; 43 out of 60 ranked CNVR were used to obtain this MSE. As with the MLR model, there was no systematic association between the number of CNVR used and prediction accuracy. The majority of these MSEs were smaller than the baseline training MSEs, for both birth and weaning weight. All minimized MSEs were smaller than the baseline training MSEs. The RF model tended to choose larger numbers of CNVR compared to the other models and fit this population better. Overall, the BSLMM and RF models yielded similar results, which were better than those from the other models. For the MLR model, we obtained some collections of CNVR having MSEs for both birth weight and weaning weight slightly smaller than the baseline training MSEs, which were similar to the MSEs from the RT model. Fitting the MLR, RT and RF models did not yield large differences on the MSEs for varying numbers of ranked CNVR for most CNVR-SNP sets, which indicate that there may be

Table V.3 Number of ranked CNVR to obtain minimum MSE for RF model.

Phenotype	CNVR-SNP set	Total number of valid CNVR analyzed	Best number of ranked CNVR	Minimum MSE	Correlation	Adjusted R^2	BIC
Birth weight	1	42 of 94	3	0.81	0.40	1.03	140.20
	2	45 of 88	16	0.82	0.36	1.01	390.83
	3	34 of 66	30	0.85	0.33	1.05	83.30
	4	72 of 149	26	0.77	0.43	1.00	1721.95
	5	73 of 121	40	0.83	0.33	1.01	409.00
	6	44 of 93	4	0.82	0.34	-0.18	1.67
Weaning weight	1	60 of 94	58	0.56	0.67	1.00	1739.09
	2	60 of 88	43	0.55	0.67	1.02	205.09
	3	37 of 66	36	0.57	0.66	1.02	193.16
	4	82 of 149	23	0.55	0.67	1.01	673.44
	5	90 of 121	27	0.57	0.66	1.04	110.14
	6	51 of 93	44	0.57	0.66	1.02	171.95

The best number of ranked CNVR to obtain minimum MSE from CV for RF model is summarized for each of the CNVR-SNP set. Their corresponding correlation, R^2 – adjusted and BIC are summarized as well. The best CNVR-SNP sets are marked in bold. Note that due to computation time, for CNVR-SNP set 4, only CNVR with MSEs less than baseline MSEs were analyzed (0.8864 for birth weight and 0.6004 for weaning weight); for other CNVR-SNP sets, CNVR with MSE less than 0.89 for birth weight and 0.61 for weaning weight were analyzed.

multiple “good” numbers of CNVR to use for prediction since MSEs for lots of them were close. And these models lack the strength to select a most appropriate number of CNVR which perform much better than all others. In addition, for the best collections of CNVR for each model, the values of model fit statistics were similar, indicating that none of these models yielded significantly better results than others.

V.3.2 Evaluating prediction accuracy of tuned models with testing data

The model fit statistics for prediction in testing data set using best collections of CNVR for each of the CNVR-SNP set are summarized in Appendix B, Table B-9 for BSLMM, RT and RF models. The model fit statistics for prediction using the best CNVR-SNP set obtained from training step for each of the three models are summarized in Table V.4, and the CNVR used in these models are summarized in Appendix C, Additional File C-27. Since the baseline prediction MSEs were 1.0110 for birth weight and 0.5672 for weaning weight, and all MSEs from BSLMM model were below them, it seems that the BSLMM model fit the data well and made good predictions. CNVR-SNP set 5 yielded minimum MSE for prediction of birth weight (0.95), and CNVR-SNP set 4 yielded minimum MSE for prediction of weaning weight (0.52).

For prediction of birth weight using RT model, only CNVR-SNP set 1 and 6 yielded MSEs slightly less than baseline prediction MSEs. For prediction of weaning weight using this model, CNVR-SNP set 3 and 4 yielded MSEs less than MSE from that MLR model. CNVR-SNP set 1 yielded minimum MSE for prediction of birth weight (1.00), and CNVR-SNP set 4 yielded minimum MSE for prediction of weaning weight (0.54).

Table V.4 Prediction in testing data set using BSLMM, RT and RF models.

Model	Birth weight				Weaning weight			
	MSE	**Correl a-tion**	**Adj. R^2**	**BIC**	**MSE**	**Correl a-tion**	**Adj. R^2**	**BIC**
BSLMM	0.98 (0.11)	0.30	1.01	571.24	0.52 (0.08)	0.68	1.01	565.76
RT	1.03 (0.06)	0.23	4.41	9.11	0.57 (0.04)	0.64	-0.45	2.82
RF	0.97 (0.00)	0.31	1.02	276.7	0.55 (0.00)	0.66	1.23	33.69

MSEs and standard deviations of MSEs (in parentheses) of prediction using the best collections of ranked CNVR in testing data set for BSLMM, RT and RF models are summarized in this table. The best collections of CNVR were chosen from CNVR-SNP sets 1-6 for each of the 3 models. Their corresponding correlation, adjusted R^2 and BIC are summarized as well.

For prediction of birth weight using RF model, CNVR-SNP set 1, 2, 3, 4 and 5 yielded MSEs less than baseline prediction MSEs. For prediction of weaning weight using RF model, CNVR-SNP set 1, 2, 4 and 5 yielded MSEs less than MSE from that MLR model. CNVR-SNP set 1 yielded minimum MSE for prediction of both birth weight (0.96) and weaning weight (0.52), which were similar to the minimum MSEs from BSLMM model by CNVR-SNP sets 5 and 4, respectively, and relatively smaller than the minimum MSEs from RT model.

The model fit statistics for prediction in testing data set using MLR model were summarized in Figure V.5. The CNVR used in these models are summarized in Appendix C, Additional File C-27. None of the CNVR collections from Appendix B, Table B-8 yielded MSEs less than baseline prediction MSEs for birth and weaning weights simultaneously.

The best CNVR-SNP set obtained from training step does not necessarily yield best MSEs for the prediction step, according to Table V.4 and Appendix B, Table B-9. Prediction accuracy was much better for weaning weight than for birth weight, since it had much lower MSE values and much higher Pearson correlation values.

V.3.3 Assessing prediction accuracy by Bootstrap

To assess prediction accuracy of the models, bootstrap was performed for each of the models using the collection of CNVR that yielded minimum MSEs for each CNVR-SNP set. The standard deviations of MSEs from bootstrap for the four models are summarized in Table V.4 and Appendix B, Table B-10.

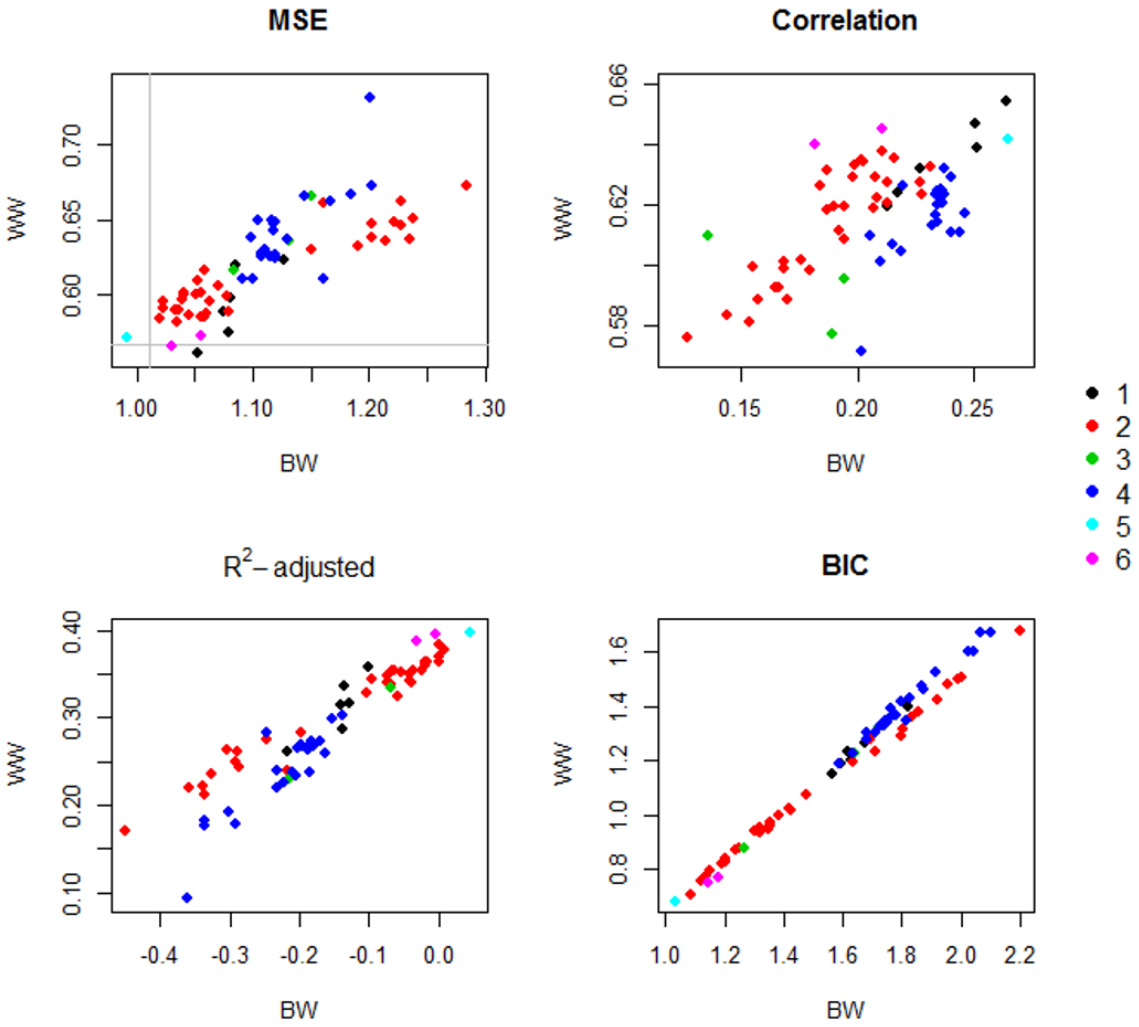


Figure V.5 Prediction in testing data set using MLR model. MSE, correlation, adjusted R^2 and BIC of the predictions in testing data set using best collections of CNVR selected by MLR model for each of the CNVR-SNP set are summarized. Different colors represent different CNVR-SNP sets. BW indicate birth weight and WW indicate weaning weight.

The BSLMM model had relatively high standard deviations for MSEs (about 20% of MSE from prediction of both birth and weaning weight), which indicate that the prediction accuracy may be poor. However, this might due to the fact that the standard deviations of MSEs for BSLMM model were obtained using a different method compared to standard bootstrap used in the other models.

The MLR and RT models had smaller standard deviations of MSEs (about 5% of MSE from prediction of birth weight and 9% of MSE from prediction of weaning weight) compared to that of BSLMM model, which indicate they have better prediction accuracy compared to BSLMM model. The standard deviations of MSEs for the RF model were the lowest and close to 0, indicating it was the model with highest prediction accuracy among the four models.

Overall, the BSLMM model yielded the smallest MSEs for prediction in testing data set for both birth and weaning weights. RF model yielded slightly larger MSEs. RT model yielded larger MSEs compared to BSLMM and RF model. MSEs of these three models were all smaller than baseline prediction MSEs. However, the MSEs we obtained are still close to each other, and the decrease of MSE after fitting the predictive models were not large, which indicates that the CNVR have some effect on prediction of birth and weaning weight for the cattle population, but the effect was small. Also, the MLR model, which was expected to yield better prediction results than the other models, turned out to be no better than the other three models. It failed to yield MSEs smaller than baseline prediction MSEs.

In addition, although the MSEs for the BSLMM model were relatively small for prediction in testing data set, compared to the other three models, the standard deviation of its MSEs were the largest, which indicates that this model was not safe and stable to use for future predictions compared to the other three models. However, the large standard deviation may be due to the different approach used in bootstrap. The other three models had relatively small standard deviations of MSEs, meaning their predictions are more stable. The RF model performs best since it had relatively good prediction performance and highest prediction accuracy.

V.3.4 Comparison of three SNP models: additive, dominant and recessive

For each of the 4 prediction models, the 3 different SNP models did not seem to yield results with large differences in MSEs and standard deviation of MSEs. That is to say, within each prediction model, the different SNP models had similar predictive performances and accuracies. In training data sets, the MSEs from additive model were slightly smaller than the other two models; in testing data sets, additive model seemed to perform slightly better than dominant model, and then better than recessive model. And although the MSEs were similar for the three SNP models, the CNVR identified to yield minimum MSEs were quite different for different SNP models using the same prediction model.

V.3.5 Genes overlapping with best collections of CNVR

The number of RefSeq genes overlapping with best collections of CNVR for each model were summarized in Table V.5. The details of the RefSeq genes overlapping with these CNVR were summarized in Appendix C, Additional File C-28. For BSLMM,

Table V.5 Number of RefSeq genes overlapping with best collections of CNVR for each model.

	CNV-SNP set	Model			
		BSLMM	**MLR**	**RT**	**RF**
BW	1	6	6	0	3
	2	6	19	3	10
	3	6	12	1	22
	4	19	22	0	2
	5	12	3	0	21
	6	5	2	0	1
	Total	38	52	4	46
WW	1	5	6	0	43
	2	7	19	1	27
	3	5	12	1	24
	4	19	22	2	1
	5	9	3	1	13
	6	3	2	1	24
	Total	35	52	5	77

The number of RefSeq genes overlapping with best collections of CNVR for each model are summarized for birth weight (BW) and weaning weight (WW). For MLR the number are the same for BW and WW since one set of CNVR were used to predict both phenotypes. The totals are without duplications.

RT and RF models, only one best collection for each CNVR-SNP set was used; for MLR model, varying numbers of lists of CNVR collections for each CNVR-SNP set, as shown in Figure V.4 (b) and Appendix B, Table B-8 were used to map to the RefSeq genes. Overall, the fewest genes were mapped by RT model; the most genes were mapped by MLR and RF models, further indicating that RT model didn't provide appropriate fit, whereas RF model provided relatively good fit to the population.

For birth weight in BSLMM model, the gene overlapping with most CNVR in the 6 CNVR-SNP data sets was LOC101907253 (T-cell receptor alpha chain V region PHDS58-like, NCBI uid 101907253) with 3 overlaps; for RF, the gene overlapping with most CNVR was RNF122 (ring finger protein 122, NCBI uid 510037) with 5 overlaps. For RT, all genes only overlapped with at most one CNVR, so they were not taken into account. For weaning weight, the gene overlapping with most CNVR was again RNF122, with 5 overlaps for BSLMM, 2 overlaps for RT, and 8 overlaps for RF. For MLR which models birth and weaning weight together, the gene overlapping with most CNVR was still RNF122, with 122 overlaps. LOC101907227 (putative protein FAM90A12P, NCBI uid 101907227) ranked the second for MLR with 67 overlaps, which also ranked second with 6 overlaps in RF for weaning weight. ULBP11 (UL16-binding protein 11, NCBI uid 510707) ranked the ninth for MLR with 19 overlaps, and ranked second with 3 overlaps in RF for birth weight. The top genes overlapping with most CNVR were summarized in Appendix C, Additional File C-29.

Two diagrams showing the common genes identified by the four models are shown in Figure V.6. For birth weight, the common genes were: RNF122, LGR6

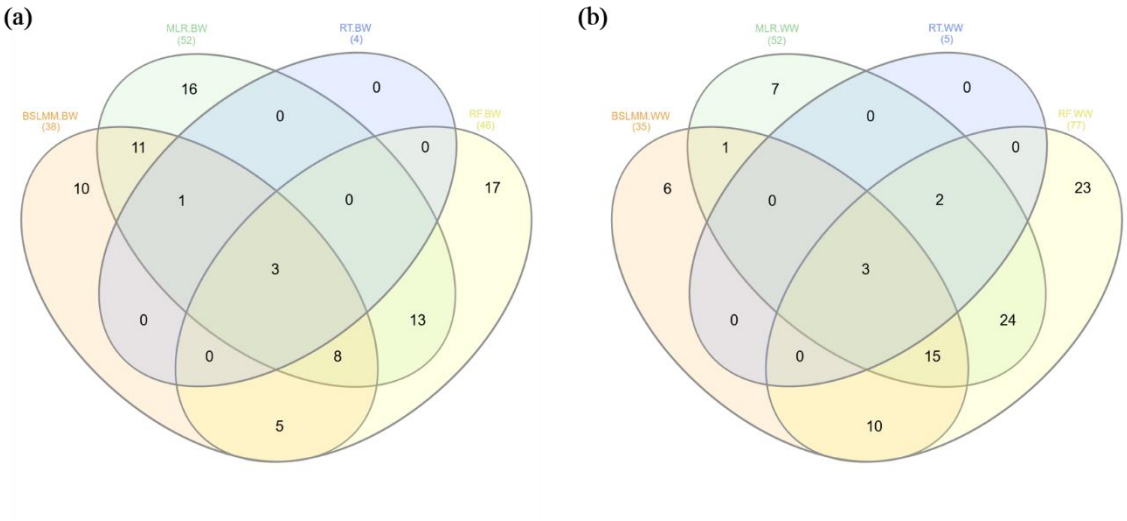


Figure V.6 The common genes identified by the four models. The common genes identified by the four models are summarized in the Venn diagrams. BW indicate birth weight and WW indicate weaning weight. The diagram was generated by InteractiVenn by Heberle et al. [98]

(leucine rich repeat containing G protein-coupled receptor 6, NCBI uid 100336662) and LOC101907253 (T-cell receptor alpha chain V region PHDS58-like, NCBI uid 101907253). For weaning weight, the common genes were: RNF122, MARK2 (microtubule affinity regulating kinase 2, NCBI uid 535197) and an unknown gene with gene ID 100296090.

V.3.6 Discussion of model performance

For BSLMM model, in Lee et al.'s study [158], genotypic data as well as the family information were both used to yield higher correlation between actual and predicted phenotypes. In our study, however, genotypic information and basic information about animals like sex and birth season were used to model the effect not explained by CNVR. Family information was not considered since its heredity was included in the genotypic information. Our correlation for best CNVR collections for each of the four models were around 0.66, which was comparable to the correlations they obtained from their data, although our sample size was less than a half of theirs. They were comparable to the correlations in Zhou et al.'s human study [160] as well, though their study had larger sample sizes (about 2 to 4 times) and number of SNP (about 2 to 4 times) compared to ours. In Ober et al.'s study [183], it was discovered that the predictive ability of SNP kept increasing until above 150,000 SNP were involved; however, in our study, the SNP with highest predictive ability were much fewer than 150000, and the ability of prediction in training data set first increase then decrease as the number of SNP increase. This difference might be due to: 1) the difference between the organisms (*Drosophila* was used in their study); 2) many SNP were close to each

other and may have high LD between them in our data; 3) we were using CNVR-SNP sets to detect the effect of CNVR on phenotypic expression, thus each CNVR was represented by a set of SNP, not individual SNP; 4) we ranked the CNVR-SNP sets first by MSEs from CV and then accumulate the number of SNP associated with ranked CNVR, therefore the SNP selected in our study may be more efficient than Ober et al.'s.

For MLR model, we used lasso shrinkage methods to overcome multicollinearity issues and tuning by v-fold CV like what Kim et al. [164] and Peng et al. [172] did in their studies. However, instead of developing new models directly, we used glmnet package in R to perform shrinkage on birth and weaning weights separately, and then used SNP that were selected by either of the two traits to fit the MLR model. But both of their studies analyzed much fewer SNP than in our study. The reason that MLR model did not yield ideal performance as expected in our study may be because of two reasons: 1) birth and weaning weights are supposed to have genetic correlation (0.5 in cattle [184] and 0.59 in swine [185]), but their correlation was weak in our data ($r^2 = 0.29$); 2) there were a number of SNP associated with each CNVR, and the SNP were close together thus might have high LD between each other. This is also the reason why lasso shrinkage need to be involved to reduce multicollinearity.

For RF model, Goldstein et al. [174] performed similar studies with different settings for the three tuning parameters: Number of variables to choose per node (mtry), Number of Trees to Grow (ntree) and weighting, and found that when mtry = 0.1 to 1 times the number of SNP and ntree is more than 250, the error rate was minimized; but weighting did not play an important role on the gain of prediction accuracy. Thus, we

randomly chose 10 CNVR form CNVR-SNP set 1 to perform 10-fold CV with various parameter settings to compare their performance, as discussed in the previous sections and Appendix A, Figure A-12. Our study finally chose default parameter settings and yielded similar MSEs compared to Heslot et al.'s [165] study in wheat.

We used four model fit statistics: MSE, Pearson correlation, $R^2 -$ adjusted and BIC. However, since number of features >> number of observations (n >> p), $R^2 -$ adjusted and BIC may be inadequate. Their weird values in our study further proved this.

Overall, in our study, there's not much difference among the additive, dominant or recessive models, but the additive model performed slightly better than the other models in most cases. For the statistical models, although RF tended to be the best model and BSLMM tended to be a good model as well, the differences of MSE and correlation values among all four models were not large. Also, MSEs from the four models didn't have large differences compared to baseline MSEs. These indicate that none of the models has a significant advantage over the other models; only subtle advantages were found. In addition, fitting these models didn't have significant advantages over fitting a MLR model with only sex, birth season and weaning age effects. These all indicate that the genetic effects of the CNVR were small. In addition, there may be several other reasons to explain this: 1) The were too many SNP in the model, resulting in n >> p, resulting in inadequate model fits; 2) Multicollinarity issues were present since the amount of SNP was large, and most SNP were close to each other which may have similar effects on the phenotype; 3) In our study, each CNVR is represented by a list of

SNP having high LD ($r^2 \geq 0.8$) with it, but not all SNP in the list represent that CNVR well and contribute significantly to the phenotype; in addition, after shrinkage and model fitting, only several (most cases less than one hundred) SNP were left in the model, which may not be a very good representation of the whole CNVR; 4) A large proportion of the SNP in our study were imputed, which may be somewhat different to the true situation and the difference is not known; 5) The values of model fit statistics are very sensitive to the setting of random seeds; 6) For MLR model, birth and weaning weight are not highly correlated. This might be due to some inaccuracies in recording the phenotype.

V.4 Conclusions

There's not much difference among additive, dominance or recessive SNP models, but additive model performed slightly better than other models in most cases. RF is the best prediction model with highest accuracy, while BSLMM is a second best model. MLR and RT models didn't yield satisfactory prediction results. However, RF and BSLMM only have subtle advantages over the other models, since the MSE and correlation values were only slightly better than the other models, and MSEs are only slightly better than baseline MSEs. Thus the effects of CNVR on birth and weaning weight are small.

In a word, the models we proposed helped phenotype prediction by CNVR to some extent, but better CNV calling methods and prediction models are waiting to be developed to better fit the population. These models could be used on other organisms including humans to predict interesting phenotypes. But one thing to keep in mind is:

since n >> p for SNP models, the best CNVR collections may not be unique. A lot of possible combinations may exist.

RNF112 and the genes we found highly associated with birth and weaning weights were not observed in other studies. Further analysis will be required to find out if these gene effects are real and how they affect the two phenotypes.

CHAPTER VI

CONCLUSION

The normalization algorithm for CNV-seq did not appear to adequately overcome large differences in depth of coverage and, consequently, the false discovery rate was grossly inflated. This issue was overcome by using different animals as the control and focusing on the common set of CNVR. Future work may show that some of the discovered CNVR contribute to variation in important phenotypes.

The performance of RD-based methods highly rely on the depth of coverage of the tested genome compared to the control genome, selection of a control animal, and selection of the window size. Larger windows achieved higher confidence for CNV detection, but resulted in higher opportunities of missing small CNV. The combination of RP and SR methods used by RAPTR-SV was the most stable and consistent among BreakDancer, CNV-seq and RAPTR-SV, with the advantage of not needing a control animal. Read pair- and SR-based methods were more stable compared to RD. However, both of them could not identify large insertions, and it was difficult for SR-based methods to identify large-scale SV. No single method was comprehensive enough, so focusing on the combined methods and the common set of CNVR could overcome the shortcomings and combine the advantages of the different detection methods.

CNVR could be tagged by SNP having high LD with them. Although there is no haplotype information for CNVR identified by WGS, CNVR which are biallelic and having odd copy numbers could be coded by a specific pattern and LD could be

calculated. SNP tagging CNVR could be used as genetic markers in GWAS and predictive modeling to identify significant SNP and CNVR, as well as important genes associated with them. Focusing just on the SNP tagging specific CNVR reduces the multiple testing problem.

In GWAS, different SNP models yielded different results; however, different SNP models had some of the same significant peaks, indicating these regions might affect the respective phenotype. Some peaks appeared at similar loci in the GWAS track of all SNP for the respective SNP model, indicating that the findings using CNVR yielded similar results to studies solely with SNP. Some peaks appeared for one or multiple CNVR-SNP sets but not with SNP alone, indicating that CNV in these regions may affect birth or weaning weights, but with SNP alone their effects did not show up because of the multiple testing problem.

The models used in this study predicted phenotypes by CNVR to some extent, but better models are waiting to be developed that better fit the population. Random forest was the best model with the highest prediction accuracy. Bayesian sparse linear mixed model also had great performance; Multivariate linear regression and regression tree didn't yield satisfactory results. The additive model had slight advantages over the dominance and recessive models. Some new genes that may have effects on birth and weaning weights in beef cattle were discovered. These models could be used on other organisms including humans to predict interesting phenotypes. Some CNVR collections were proposed to have best prediction effects, but for SNP models which have way more features than observations, a lot of other possible CNVR collections with good

performance may exist. Further analysis will be required to find out if the gene effects we discovered are real and how they affect birth and weaning weights in beef cattle.

REFERENCES

1. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nature reviews Genetics* 2006, **7**(2):85-97.
2. Sebat J: **Major changes in our DNA lead to major changes in our thinking**. *Nat Genet* 2007, **39**(7 Suppl):S3-5.
3. 1000 Genomes Project Consortium, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
4. Lupski JR: **Genomic rearrangements and sporadic disease**. *Nat Genet* 2007, **39**(7 Suppl):S43-47.
5. Tremmel R, Klein K, Winter S, Schaeffeler E, Zanger UM: **Gene copy number variation analysis reveals dosage-insensitive expression of CYP2E1**. *The pharmacogenomics journal* 2015.
6. Kim GJ, Sock E, Buchberger A, Just W, Denzer F, Hoepffner W, German J, Cole T, Mann J, Seguin JH *et al*: **Copy number variation of two separate regulatory regions upstream of SOX9 causes isolated 46,XY or 46,XX disorder of sex development**. *Journal of medical genetics* 2015, **52**(4):240-247.
7. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Stahl T, Menzel U, Sandgren J, von Tell D, Poplawski A *et al*: **Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles**. *Am J Hum Genet* 2008, **82**(3):763-771.

8. Kondrashov AS: **Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases**. *Hum Mutat* 2003, **21**(1):12-27.

9. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P *et al*: **Origins and functional impact of copy number variation in the human genome**. *Nature* 2010, **464**(7289):704-712.

10. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.

11. Cooper GM, Nickerson DA, Eichler EE: **Mutational and selective effects on copy-number variants in the human genome**. *Nat Genet* 2007, **39**(7 Suppl):S22-29.

12. Hastings PJ, Lupski JR, Rosenberg SM, Ira G: **Mechanisms of change in gene copy number**. *Nature reviews Genetics* 2009, **10**(8):551-564.

13. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C *et al*: **Relative impact of nucleotide and copy number variation on gene expression phenotypes**. *Science* 2007, **315**(5813):848-853.

14. Henrichsen CN, Chaignat E, Reymond A: **Copy number variants, diseases and gene expression**. *Hum Mol Genet* 2009, **18**(R1):R1-8.

15. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA: **Large multiallelic copy number variations in humans**. *Nat Genet* 2015, **47**(3):296-303.

16. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS *et al*: **Development and characterization of a high density SNP genotyping assay for cattle**. *PLoS One* 2009, **4**(4):e5350.

17. The Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM *et al*: **Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds**. *Science* 2009, **324**(5926):528-532.

18. Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun JY, Kim JY, Pasaje CF, Lee JS, Shin HD: **Identification of copy number variations and common deletion polymorphisms in cattle**. *BMC Genomics* 2010, **11**:232.

19. Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P, Nardone A: **Massive screening of copy number population-scale variation in Bos taurus genome**. *BMC Genomics* 2013, **14**:124.

20. Fadista J, Thomsen B, Holm LE, Bendixen C: **Copy number variation in the bovine genome**. *BMC Genomics* 2010, **11**:284.

21. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim ES, Matukumalli LK, Ventura M, Song J, VanRaden PM *et al*: **Genomic characteristics of cattle copy number variations**. *BMC Genomics* 2011, **12**:127.

22. Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, Sonstegard TS, Van Tassell CP, Liu GE: **Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins**. *BMC genomics* 2014, **15**:683.

23. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME *et al*: **Analysis of copy number variations among diverse cattle breeds**. *Genome Res* 2010, **20**(5):693-703.

24. Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S, Whan V: **Analysis of copy number variants in the cattle genome**. *Gene* 2011, **482**(1-2):73-77.

25. Stothard P, Choi JW, Basu U, Sumner-Thomson JM, Meng Y, Liao X, Moore SS: **Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery**. *BMC genomics* 2011, **12**:559.

26. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF *et al*: **Copy number variation of individual cattle genomes using next-generation sequencing**. *Genome Res* 2012, **22**(4):778-790.

27. Choi JW, Lee KT, Liao X, Stothard P, An HS, Ahn S, Lee S, Lee SY, Moore SS, Kim TH: **Genome-wide copy number variation in Hanwoo, Black Angus, and Holstein cattle**. *Mammalian genome : official journal of the International Mammalian Genome Society* 2013, **24**(3-4):151-163.

28. Choi JW, Liao X, Stothard P, Chung WH, Jeon HJ, Miller SP, Choi SY, Lee JK, Yang B, Lee KT *et al*: **Whole-genome analyses of Korean native and Holstein cattle breeds by massively parallel sequencing**. *PLoS One* 2014, **9**(7):e101127.

29. Shin DH, Lee HJ, Cho S, Kim HJ, Hwang JY, Lee CK, Jeong J, Yoon D, Kim H: **Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level**. *BMC Genomics* 2014, **15**(1):240.

30. Xie C, Tammi MT: **CNV-seq, a new method to detect copy number variation using high-throughput sequencing**. *BMC bioinformatics* 2009, **10**:80.

31. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O *et al*: **Personalized copy number and segmental duplication maps using next-generation sequencing**. *Nat Genet* 2009, **41**(10):1061-1067.

32. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A: **Statistical challenges associated with detecting copy number variations with next-generation sequencing**. *Bioinformatics* 2012, **28**(21):2711-2718.

33. Handsaker RE, Korn JM, Nemesh J, McCarroll SA: **Discovery and genotyping of genome structural polymorphism by sequencing on a population scale**. *Nat Genet* 2011, **43**(3):269-276.

34. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK *et al*: **Mapping copy number variation by population-scale genome sequencing**. *Nature* 2011, **470**(7332):59-65.

35. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M *et al*: **An integrated map of structural variation in 2,504 human genomes**. *Nature* 2015, **526**(7571):75-81.

36. Xu Y, Zhang L, Shi T, Zhou Y, Cai H, Lan X, Zhang C, Lei C, Chen H: **Copy number variations of MICAL-L2 shaping gene expression contribute to different phenotypes of cattle**. *Mammalian genome : official journal of the International Mammalian Genome Society* 2013, **24**(11-12):508-516.

37. Durkin K, Coppieters W, Drogemuller C, Ahariz N, Cambisano N, Druet T, Fasquelle C, Haile A, Horin P, Huang L *et al*: **Serial translocation by means of circular intermediates underlies colour sidedness in cattle**. *Nature* 2012, **482**(7383):81-84.

38. Venhoranta H, Pausch H, Wysocki M, Szczerbal I, Hanninen R, Taponen J, Uimari P, Flisikowski K, Lohi H, Fries R *et al*: **Ectopic KIT copy number variation underlies impaired migration of primordial germ cells associated with gonadal hypoplasia in cattle (Bos taurus)**. *PLoS One* 2013, **8**(9):e75659.

39. Zhang L, Jia S, Yang M, Xu Y, Li C, Sun J, Huang Y, Lan X, Lei C, Zhou Y *et al*: **Detection of copy number variations and their effects in Chinese bulls**. *BMC genomics* 2014, **15**:480.

40. McDaneld TG, Kuehn LA, Thomas MG, Pollak EJ, Keele JW: **Deletion on chromosome 5 associated with decreased reproductive efficiency in female cattle**. *J Anim Sci* 2014, **92**(4):1378-1384.

41. Hulsman Hanna LL, Garrick DJ, Gill CA, Herring AD, Riggs PK, Miller RK, Sanders JO, Riley DG: **Genome-wide association study of temperament and tenderness using different Bayesian approaches in a Nellore-Angus crossbred population**. *Livestock Science* 2013.

42. Aronesty E: **ea-utils: command line tools for processing biological sequencing data. http://code.google.com/p/ea-utils**. 2011.

43. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS *et al*: **A whole-genome assembly of the domestic cow, Bos taurus**. *Genome Biol* 2009, **10**(4):R42.

44. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

45. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.

46. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M: **Detecting copy number variation with mated short reads**. *Genome Res* 2010, **20**(11):1613-1622.

47. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP *et al*: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation**. *Nature methods* 2009, **6**(9):677-681.

48. Fan X, Abbott TE, Larson D, Chen K: **BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping**. *Curr Protoc Bioinformatics* 2014, **2014**.

49. Suzuki T, Tsurusaki Y, Nakashima M, Miyake N, Saitsu H, Takeda S, Matsumoto N: **Precise detection of chromosomal translocation or inversion breakpoints by whole-genome sequencing**. *Journal of human genetics* 2014.

50. Abel HJ, Al-Kateb H, Cottrell CE, Bredemeyer AJ, Pritchard CC, Grossmann AH, Wallander ML, Pfeifer JD, Lockwood CM, Duncavage EJ: **Detection of gene rearrangements in targeted clinical next-generation sequencing**. *The Journal of molecular diagnostics : JMD* 2014, **16**(4):405-417.

51. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives**. *BMC Bioinformatics* 2013, **14 Suppl 11**:S1.

52. Bickhart DM, Hutchison JL, Xu L, Schnabel RD, Taylor JF, Reecy JM, Schroeder S, Van Tassell CP, Sonstegard TS, Liu GE: **RAPTR-SV: a hybrid method for the detection of structural variants**. *Bioinformatics* 2015, **31**(13):2084-2090.

53. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: **Sensitive and accurate detection of copy number variants using read depth of coverage**. *Genome research* 2009, **19**(9):1586-1592.

54. Xi R, Kim TM, Park PJ: **Detecting structural variations in the human genome using next generation sequencing**. *Briefings in functional genomics* 2010, **9**(5-6):405-415.

55. Huang WD, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.

56. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J Roy Statist Soc Ser B* 1995, **57**(1):289-300.

57. Hou Y, Bickhart DM, Chung H, Hutchison JL, Norman HD, Connor EE, Liu GE: **Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake**. *Functional & integrative genomics* 2012, **12**(4):717-723.

58. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.

59. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies**. *Nat Genet* 2012, **44**(7):821-824.

60. Fernando RL, Garrick D: **Bayesian methods applied in GWAS**. In: *Genome-wide association studies and genomic prediction.* Edited by Gondro C, van der

Werf J, Hayes B, vol. 1019: Springer Science+Business Media, LLC; 2013: 237-274.

61. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, de Ståhl TD, Menzel U, Sandgren J, von Tell D, Poplawski A: **Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles**. *The American Journal of Human Genetics* 2008, **82**(3):763-771.

62. Lupski JR: **Genomic rearrangements and sporadic disease**. *Nature genetics* 2007, **39**:S43.

63. Kondrashov AS: **Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases**. *Human mutation* 2003, **21**(1):12-27.

64. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P: **Origins and functional impact of copy number variation in the human genome**. *Nature* 2010, **464**(7289):704.

65. Consortium GP: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68.

66. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444.

67. Cooper GM, Nickerson DA, Eichler EE: **Mutational and selective effects on copy-number variants in the human genome**. *Nature genetics* 2007, **39**:S22.

68. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, De Grassi A, Lee C: **Relative impact of nucleotide and copy number variation on gene expression phenotypes**. *Science* 2007, **315**(5813):848-853.

69. Henrichsen CN, Chaignat E, Reymond A: **Copy number variants, diseases and gene expression**. *Human molecular genetics* 2009, **18**(R1):R1-R8.

70. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA: **Large multiallelic copy number variations in humans**. *Nature genetics* 2015, **47**(3):296.

71. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'connell J, Moore SS, Smith TP, Sonstegard TS: **Development and characterization of a high density SNP genotyping assay for cattle**. *PloS one* 2009, **4**(4):e5350.

72. Consortium BH: **Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds**. *Science* 2009, **324**(5926):528-532.

73. Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun J-Y, Kim JY, Pasaje CFA, Lee JS, Shin HD: **Identification of copy number variations and common deletion polymorphisms in cattle**. *BMC genomics* 2010, **11**(1):232.

74. Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P, Nardone A: **Massive screening of copy number population-scale variation in Bos taurus genome**. *BMC genomics* 2013, **14**(1):124.

75. Fadista J, Thomsen B, Holm L-E, Bendixen C: **Copy number variation in the bovine genome**. *BMC genomics* 2010, **11**(1):284.

76. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim E-s, Matukumalli LK, Ventura M, Song J, VanRaden PM: **Genomic characteristics of cattle copy number variations**. *BMC genomics* 2011, **12**(1):127.

77. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME: **Analysis of copy number variations among diverse cattle breeds**. *Genome research* 2010.

78. Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S, Whan V: **Analysis of copy number variants in the cattle genome**. *Gene* 2011, **482**(1):73-77.

79. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF *et al*: **Copy number variation of individual cattle genomes using next-generation sequencing**. *Genome research* 2012, **22**(4):778-790.

80. Choi J-W, Liao X, Stothard P, Chung W-H, Jeon H-J, Miller SP, Choi S-Y, Lee J-K, Yang B, Lee K-T: **Whole-genome analyses of Korean native and Holstein cattle breeds by massively parallel sequencing**. *PloS one* 2014, **9**(7):e101127.

81. Shin DH, Lee HJ, Cho S, Kim HJ, Hwang JY, Lee CK, Jeong J, Yoon D, Kim H: **Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level**. *BMC genomics* 2014, **15**:240.

82. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O: **Personalized copy number and**

segmental duplication maps using next-generation sequencing. *Nature genetics* 2009, **41**(10):1061.

83. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC: **mrsFAST: a cache-oblivious algorithm for short-read mapping**. *Nature methods* 2010, **7**(8):576.

84. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A: **Statistical challenges associated with detecting copy number variations with next-generation sequencing**. *Bioinformatics* 2012, **28**(21):2711-2718.

85. Riley DG, Welsh TH, Gill CA, Hulsman LL, Herring AD, Riggs PK, Sawyer JE, Sanders JO: **Whole genome association of SNP with newborn calf cannon bone length**. *Livestock Science* 2013, **155**(2-3):186-196.

86. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic acids research* 2009, **38**(6):1767-1771.

87. Aronesty E: **ea-utils: Command-line tools for processing biological sequencing data**. *Durham, NC: Expression Analysis* 2011.

88. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS: **A whole-genome assembly of the domestic cow, Bos taurus**. *Genome biology* 2009, **10**(4):R42.

89. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

90. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome research* 2010.

91. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

92. Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C: **Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping**. *BMC genomics* 2011, **12**:557.

93. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2008, **4**(1):44.

94. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA: **Primer3Plus, an enhanced web interface to Primer3**. *Nucleic acids research* 2007, **35**(Web Server issue):W71-74.

95. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method**. *methods* 2001, **25**(4):402-408.

96. Weaver S, Dube S, Mir A, Qin J, Sun G, Ramakrishnan R, Jones RC, Livak KJ: **Taking qPCR to a higher level: analysis of CNV reveals the power of high**

throughput qPCR to enhance quantitative resolution. *Methods* 2010, **50**(4):271-276.

97. Janevski A, Varadan V, Kamalakaran S, Banerjee N, Dimitrova N: **Effective normalization for copy number variation detection from whole genome sequencing**. *BMC genomics* 2012, **13**(6):S16.

98. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R: **InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams**. *BMC bioinformatics* 2015, **16**:169.

99. Nozawa M, Kawahara Y, Nei M: **Genomic drift and copy number variation of sensory receptor genes in humans**. *Proceedings of the National Academy of Sciences* 2007, **104**(51):20421-20426.

100. Paudel Y, Madsen O, Megens H-J, Frantz LA, Bosse M, Crooijmans RP, Groenen MA: **Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors**. *BMC genomics* 2015, **16**(1):330.

101. Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, Ezra E, Zeron Y: **Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs**. *BMC genomics* 2010, **11**(1):673.

102. Wu Y, Fan H, Jing S, Xia J, Chen Y, Zhang L, Gao X, Li J, Gao H, Ren H: **A genome-wide scan for copy number variations using high-density single nucleotide polymorphism array in Simmental cattle**. *Animal genetics* 2015, **46**(3):289-298.

103. Xu L, Hou Y, Bickhart DM, Song J, Van Tassell CP, Sonstegard TS, Liu GE: **A genome-wide survey reveals a deletion polymorphism associated with resistance to gastrointestinal nematodes in Angus cattle**. *Functional & integrative genomics* 2014, **14**(2):333-339.

104. Hu Z-L, Fritz ER, Reecy JM: **AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond**. *Nucleic acids research* 2006, **35**(suppl_1):D604-D609.

105. Buysse K, Delle Chiaie B, Van Coster R, Loeys B, De Paepe A, Mortier G, Speleman F, Menten B: **Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience**. *European journal of medical genetics* 2009, **52**(6):398-403.

106. Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays**. *Nature genetics* 2007, **39**:S16.

107. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K: **Assembly of microarrays for genome-wide measurement of DNA copy number**. *Nature genetics* 2001, **29**(3):263.

108. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nature biotechnology* 2008, **26**(10):1135.

109. Pirooznia M, Goes FS, Zandi PP: **Whole-genome CNV analysis: advances in computational approaches**. *Frontiers in genetics* 2015, **6**:138.

110. Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M: **Read count approach for DNA copy number variants detection**. *Bioinformatics* 2011, **28**(4):470-478.

111. Xi R, Luquette J, Hadjipanayis A, Kim T-M, Park PJ: **BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data**. *Genome biology* 2010, **11**(S1):O10.

112. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing**. *Genome research* 2011:gr. 114876.114110.

113. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L *et al*: **Paired-end mapping reveals extensive structural variation in the human genome**. *Science* 2007, **318**(5849):420-426.

114. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives**. *BMC bioinformatics* 2013, **14**(11):S1.

115. Medvedev P, Stanciu M, Brudno M: **Computational methods for discovering structural variation with next-generation sequencing**. *Nature methods* 2009, **6**(11 Suppl):S13-20.

116. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB: **PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data**. *Genome biology* 2009, **10**(2):R23.

117. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, Gerstein M: **Identification of genomic indels and structural variations using split reads**. *BMC genomics* 2011, **12**:375.

118. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: **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):2865-2871.

119. Abyzov A, Gerstein M: **AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision**. *Bioinformatics* 2011, **27**(5):595-603.

120. Nijkamp JF, van den Broek MA, Geertman J-MA, Reinders MJ, Daran J-MG, de Ridder D: **De novo detection of copy number variation by co-assembly**. *Bioinformatics* 2012, **28**(24):3195-3202.

121. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs**. *Nature genetics* 2012, **44**(2):226.

122. Norris BJ, Whan VA: **A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep**. *Genome research* 2008:gr. 072090.072107.

123. Fontanesi L, Beretti F, Riggio V, González EG, Dall'Olio S, Davoli R, Russo V, Portolano B: **Copy number variation and missense mutations of the agouti**

signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenetic and genome research* 2009, **126**(4):333-347.

124. Le Scouarnec S, Gribble SM: **Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics**. *Heredity* 2012, **108**(1):75-85.

125. Alkan C, Coe BP, Eichler EE: **Genome structural variation discovery and genotyping**. *Nature reviews Genetics* 2011, **12**(5):363-376.

126. Xu L, Hou Y, Bickhart DM, Zhou Y, Song J, Sonstegard TS, Van Tassell CP, Liu GE: **Population-genetic properties of differentiated copy number variations in cattle**. *Scientific reports* 2016, **6**:23161.

127. Zanda M, Onengut-Gumuscu S, Walker N, Shtir C, Gallo D, Wallace C, Smyth D, Todd JA, Hurles ME, Plagnol V: **A genome-wide assessment of the role of untagged copy number variants in type 1 diabetes**. *PLoS genetics* 2014, **10**(5):e1004367.

128. Wangler MF, Hu Y, Shulman JM: **Drosophila and genome-wide association studies: a review and resource for the functional dissection of human complex traits**. *Disease models & mechanisms* 2017, **10**(2):77-88.

129. Fortes MR, Reverter A, Zhang Y, Collis E, Nagaraj SH, Jonsson NN, Prayaga KC, Barris W, Hawken RJ: **Association weight matrix for the genetic dissection of puberty in beef cattle**. *Proceedings of the National Academy of Sciences* 2010, **107**(31):13642-13647.

130. Kadri NK, Koks PD, Meuwissen TH: **Prediction of a deletion copy number variant by a dense SNP panel**. *Genetics, selection, evolution : GSE* 2012, **44**:7.

131. McCarroll S, N Hadnott T, Perry G, Sabeti P, C Zody M, C Barrett J, Dallaire S, B Gabriel S, Lee C, Daly M *et al*: **The International HapMap Consortium Common deletion polymorphisms in the human genome**, vol. 38; 2006.

132. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *American journal of human genetics* 2007, **81**(3):559-575.

133. **PLINK v1.90b3.45** [http://pngu.mgh.harvard.edu/purcell/plink/]

134. Zhou X: **A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies**. *bioRxiv* 2016.

135. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies**. *Nature genetics* 2012, **44**(7):821-824.

136. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency**. *Annals of statistics* 2001:1165-1188.

137. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the royal statistical society Series B (Methodological)* 1995:289-300.

138. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics**. *Genome research* 2009.

139. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ: **Biological, clinical and population relevance of 95 loci for blood lipids**. *Nature* 2010, **466**(7307):707.

140. Shim H, Chasman DI, Smith JD, Mora S, Ridker PM, Nickerson DA, Krauss RM, Stephens M: **A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians**. *PloS one* 2015, **10**(4):e0120758.

141. Neupane M, Kiser J, Team BRDCCAPR, Neibergs H: **Gene set enrichment analysis of SNP data in dairy and beef cattle with bovine respiratory disease**. *Animal genetics* 2018.

142. Ghosh S, Ghosh A, Maiti GP, Mukherjee N, Dutta S, Roy A, Roychoudhury S, Panda CK: **LIMD1 is more frequently altered than RB1 in head and neck squamous cell carcinoma: clinical and prognostic implications**. *Molecular cancer* 2010, **9**(1):58.

143. Anton I, Húth B, Füller I, Gábor G, Holló G, Zsolnai A: **Effect of single-nucleotide polymorphisms on the breeding value of fertility and breeding value of beef in Hungarian Simmental cattle**. *Acta Veterinaria Hungarica* 2018, **66**(2):215-225.

144. Bernaciak J, Wiśniowiecka-Kowalnik B, Castañeda J, Kutkowska-Kaźmierczak A, Nowakowska B: **A NOVEL DE NOVO 20q13. 11q13. 12 MICRODELETION IN A BOY WITH NEURODEVELOPMENTAL DISORDERS**.

145. Kushima I, Aleksic B, Nakatochi M, Shimamura T, Okada T, Uno Y, Morikawa M, Ishizuka K, Shiino T, Kimura H: **Comparative Analyses of Copy-Number Variation in Autism Spectrum Disorder and Schizophrenia Reveal Etiological Overlap and Biological Insights**. *Cell reports* 2018, **24**(11):2838-2856.

146. Chang X, Xia Y, Pan J, Meng Q, Zhao Y, Yan X: **PADI2 is significantly associated with rheumatoid arthritis**. *PloS one* 2013, **8**(12):e81259.

147. Khajavi M, Zhou Y, Birsner AE, Bazinet L, Di Sant AR, Schiffer AJ, Rogers MS, Krishnaji ST, Hu B, Nguyen V: **Identification of Padi2 as a novel angiogenesis-regulating gene by genome association studies in mice**. *PLoS genetics* 2017, **13**(6):e1006848.

148. Shiels A, Bennett TM, Knopf HL, Maraini G, Li A, Jiao X, Hejtmancik JF: **The EPHA2 gene is associated with cataracts linked to chromosome 1p**. *Molecular vision* 2008, **14**:2042.

149. Sundaresan P, Ravindran RD, Vashist P, Shanker A, Nitsch D, Talwar B, Maraini G, Camparini M, Nonyane BAS, Smeeth L: **EPHA2 polymorphisms and age-related cataract in India**. *PloS one* 2012, **7**(3):e33001.

150. Tavira B, Gómez J, Ortega F, Tranche S, Díaz-Corte C, Alvarez F, Ortiz A, Santos F, Sánchez-Niño MD, Coto E: **A CLCNKA polymorphism (rs10927887; p. Arg83Gly) previously linked to heart failure is associated with the estimated glomerular filtration rate in the RENASTUR cohort**. *Gene* 2013, **527**(2):670-672.

151. Stark K, Esslinger UB, Reinhard W, Petrov G, Winkler T, Komajda M, Isnard R, Charron P, Villard E, Cambien F: **Genetic association study identifies HSPB7 as a risk gene for idiopathic dilated cardiomyopathy**. *PLoS genetics* 2010, **6**(10):e1001167.

152. Armengol L, Gratacos M, Pujana M, Ribasés M, Martin-Santos R, Estivill X: **5′ UTR-region SNP in the NTRK3 gene is associated with panic disorder**. *Molecular psychiatry* 2002, **7**(9):928.

153. Athanasiu L, Mattingsdal M, Melle I, Inderhaug E, Lien T, Agartz I, Lorentzen S, Morken G, Andreassen OA, Djurovic S: **Intron 12 in NTRK3 is associated with bipolar disorder**. *Psychiatry research* 2011, **185**(3):358-362.

154. Muiños-Gimeno M, Guidi M, Kagerbauer B, Martín-Santos R, Navinés R, Alonso P, Menchón JM, Gratacòs M, Estivill X, Espinosa-Parrilla Y: **Allele variants in functional microRNA target sites of the neurotrophin-3 receptor gene (NTRK3) as susceptibility factors for anxiety disorders**. *Human mutation* 2009, **30**(7):1062-1071.

155. Mercader JM, Saus E, Agüera Z, Bayés M, Boni C, Carreras A, Cellini E, De Cid R, Dierssen M, Escaramís G: **Association of NTRK3 and its interaction with NGF suggest an altered cross-regulation of the neurotrophin signaling pathway in eating disorders**. *Human molecular genetics* 2008, **17**(9):1234-1244.

156. Verma R, Holmans P, Knowles JA, Grover D, Evgrafov OV, Crowe RR, Scheftner WA, Weissman MM, DePaulo Jr JR, Potash JB: **Linkage**

disequilibrium mapping of a chromosome 15q25-26 major depression linkage region and sequencing of NTRK3. *Biological psychiatry* 2008, **63**(12):1185-1189.

157. Otnæss MK, Djurovic S, Rimol LM, Kulle B, Kähler AK, Jönsson EG, Agartz I, Sundet K, Hall H, Timm S: **Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia**. *Neurobiology of disease* 2009, **34**(3):518-524.

158. Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM: **Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data**. *PLoS genetics* 2008, **4**(10):e1000231.

159. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM: **Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model**. *PLoS genetics* 2015, **11**(4):e1004969.

160. Zhou X, Carbonetto P, Stephens M: **Polygenic modeling with bayesian sparse linear mixed models**. *PLoS genetics* 2013, **9**(2):e1003264.

161. Johnson RA, Wichern DW: **Applied multivariate statistical analysis**, vol. 4: Prentice-Hall New Jersey; 2014.

162. Knott SA, Haley CS: **Multitrait least squares for quantitative trait loci detection**. *Genetics* 2000, **156**(2):899-911.

163. Xu C, Li Z, Xu S: **Joint mapping of quantitative trait loci for multiple binary characters**. *Genetics* 2005, **169**(2):1045-1059.

164. Kim S, Sohn K-A, Xing EP: **A multivariate regression approach to association analysis of a quantitative trait network**. *Bioinformatics* 2009, **25**(12):i204-i212.

165. Heslot N, Yang H-P, Sorrells ME, Jannink J-L: **Genomic selection in plant breeding: a comparison of models**. *Crop Science* 2012, **52**(1):146-160.

166. Breiman L, Friedman J, Stone CJ, Olshen RA: **Classification and regression trees**: CRC press; 1984.

167. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.

168. Breiman L: **Random forests. Mach Learn45: 5–32**. In.; 2001.

169. Winham SJ, Colby CL, Freimuth RR, Wang X, De Andrade M, Huebner M, Biernacka JM: **SNP interaction detection with random forests in high-dimensional genetic data**. *BMC bioinformatics* 2012, **13**(1):164.

170. Zeng P, Zhou X, Huang S: **Prediction of gene expression with cis-SNPs using mixed models and regularization methods**. *BMC genomics* 2017, **18**(1):368.

171. Comings DE, Gade-Andavolu R, Gonzalez N, Wu S, Muhleman D, Blake H, Dietz G, Saucier G, P MacMurray J: **Comparison of the role of dopamine, serotonin, and noradrenaline genes in ADHD, ODD and conduct disorder: multivariate regression analysis of 20 genes**. *Clinical genetics* 2000, **57**(3):178-196.

172. Peng J, Zhu J, Bergamaschi A, Han W, Noh D-Y, Pollack JR, Wang P: **Regularized multivariate regression for identifying master predictors with**

application to integrative genomics study of breast cancer. *The annals of applied statistics* 2010, **4**(1):53.

173. Goldstein BA, Polley EC, Briggs F: **Random forests for genetic association studies**. *Statistical applications in genetics and molecular biology* 2011, **10**(1).

174. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF: **An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings**. *BMC genetics* 2010, **11**(1):49.

175. García-Magariños M, López-de-Ullibarri I, Cao R, Salas A: **Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of SNP-SNP Interaction**. *Annals of human genetics* 2009, **73**(3):360-369.

176. Schaid DJ: **General score tests for associations of genetic markers with disease using cases and their parents**. *Genetic epidemiology* 1996, **13**(5):423-449.

177. James G, Witten D, Hastie T, Tibshirani R: **An introduction to statistical learning**, vol. 112: Springer; 2013.

178. Efron B, Tibshirani RJ: **An introduction to the bootstrap**: CRC press; 1994.

179. Team RC: **R: A language and environment for statistical computing**. *R Foundation for Statistical Computing, Viena, Austria, Vienna, Austria* 2016.

180. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent**. *Journal of statistical software* 2010, **33**(1):1-22.

181. Ripley B: **tree: Classification and Regression Trees**. 2018.

182. Liaw A, Wiener M: **Classification and regression by randomForest**. *R news* 2002, **2**(3):18-22.

183. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, Stricker C, Gianola D, Schlather M, Mackay TFC *et al*: **Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster**. *PLoS genetics* 2012, **8**(5):e1002685.

184. Scott P. Greiner PD, Extension Animal Scientist, Beef, VA Tech: **Genetic Relationships**. *Virginia Cooperative Extension* 2018.

185. Kaufmann D, Hofer A, Bidanel J, Künzi N: **Genetic parameters for individual birth and weaning weight and for litter size of Large White pigs**. *Journal of Animal Breeding and Genetics* 2000, **117**(3):121-128.

APPENDIX A

FIGURES

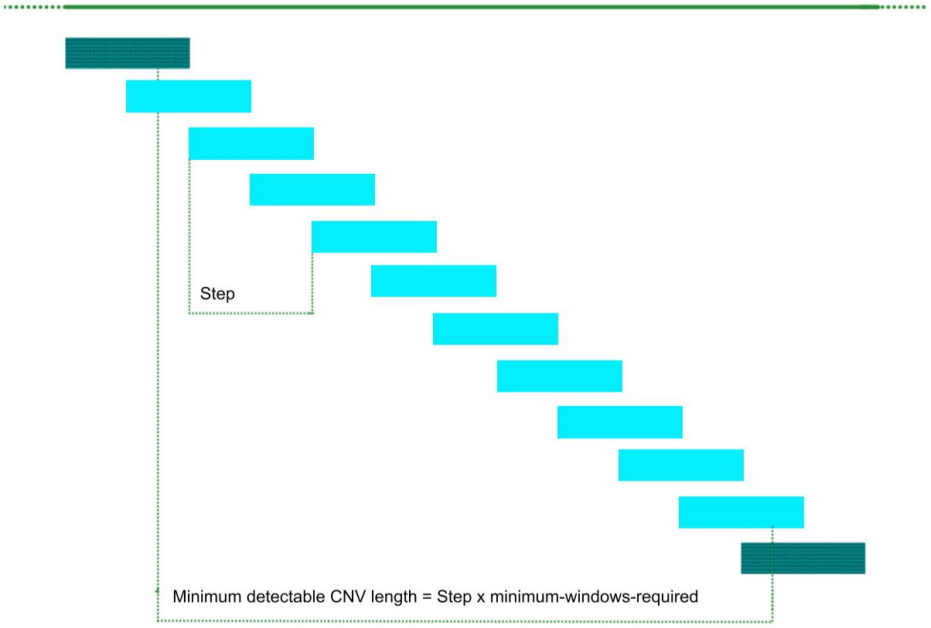


Figure A-1 Determining the length of a CNVR using cnv-seq. Cnv-seq divides the sequence of a chromosome (green line) into overlapping windows where the size of the window = --window-size * --bigger-window. The offset or step for each overlap is half the size of the window. For each window, cnv-seq determines if the |log2| ratio of sequence counts in the test animal compared to the control animal significantly exceeds the threshold log2 value. If the |log2| value does not exceed the threshold log2 value, the position of the window is not recorded (hatched boxes). If a window does significantly exceed the |log2| value, the position of the window is recorded (teal box). A CNV is annotated in the final output if the number of consecutive overlapping windows exceeding the log2 threshold is greater than or equal to the specified minimum number of windows required (--minimum-windows-required). In this example, 10 consecutive windows were required for annotation. Thus, the minimum detectable length of a CNVR is step * --minimum-windows-required. The length of a CNVR is the number of consecutive windows deviating significantly from the log2 threshold multiplied by the size of the step.

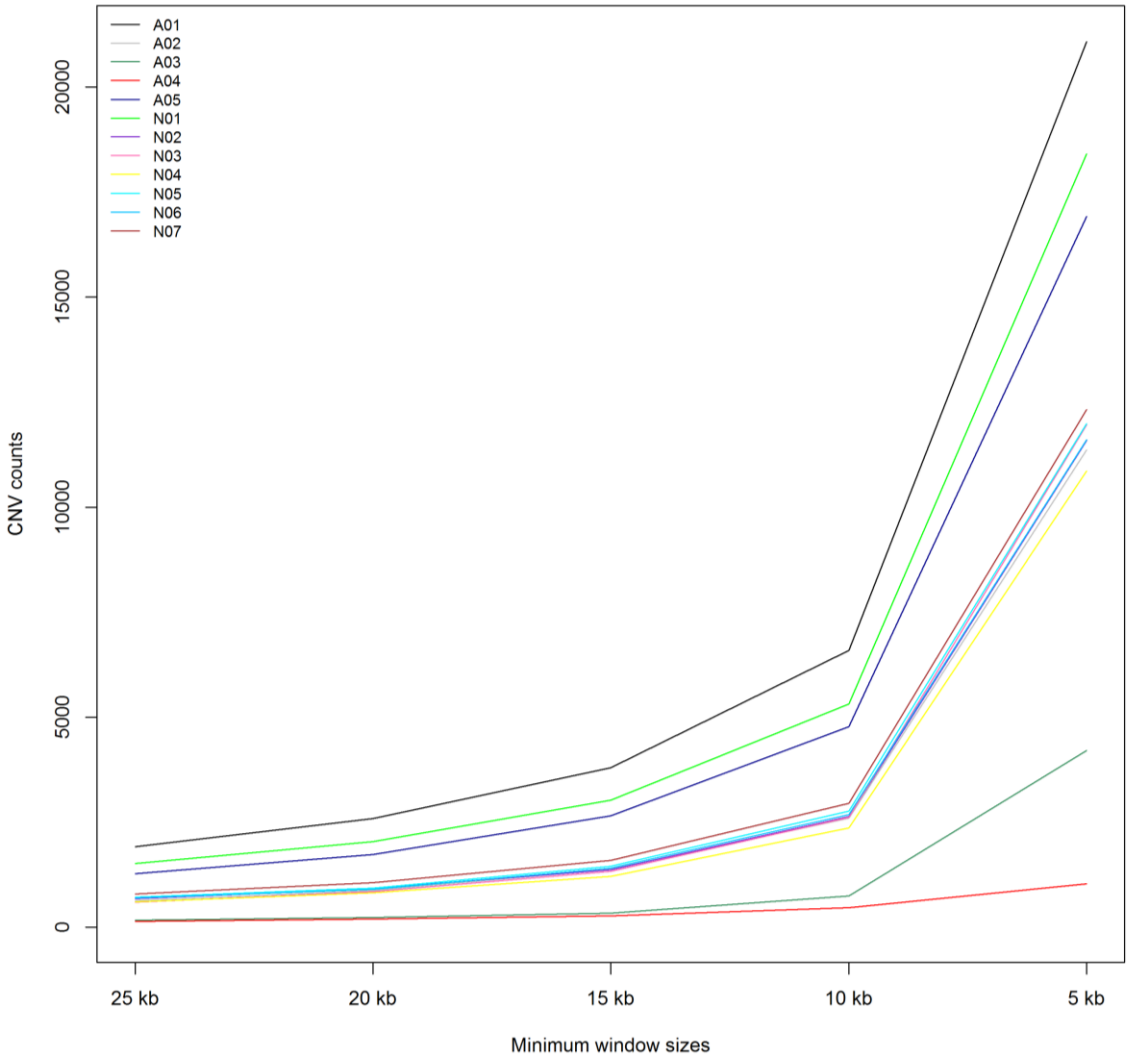


Figure A-2 Relationship of CNV counts to minimum detectable CNVR sizes.

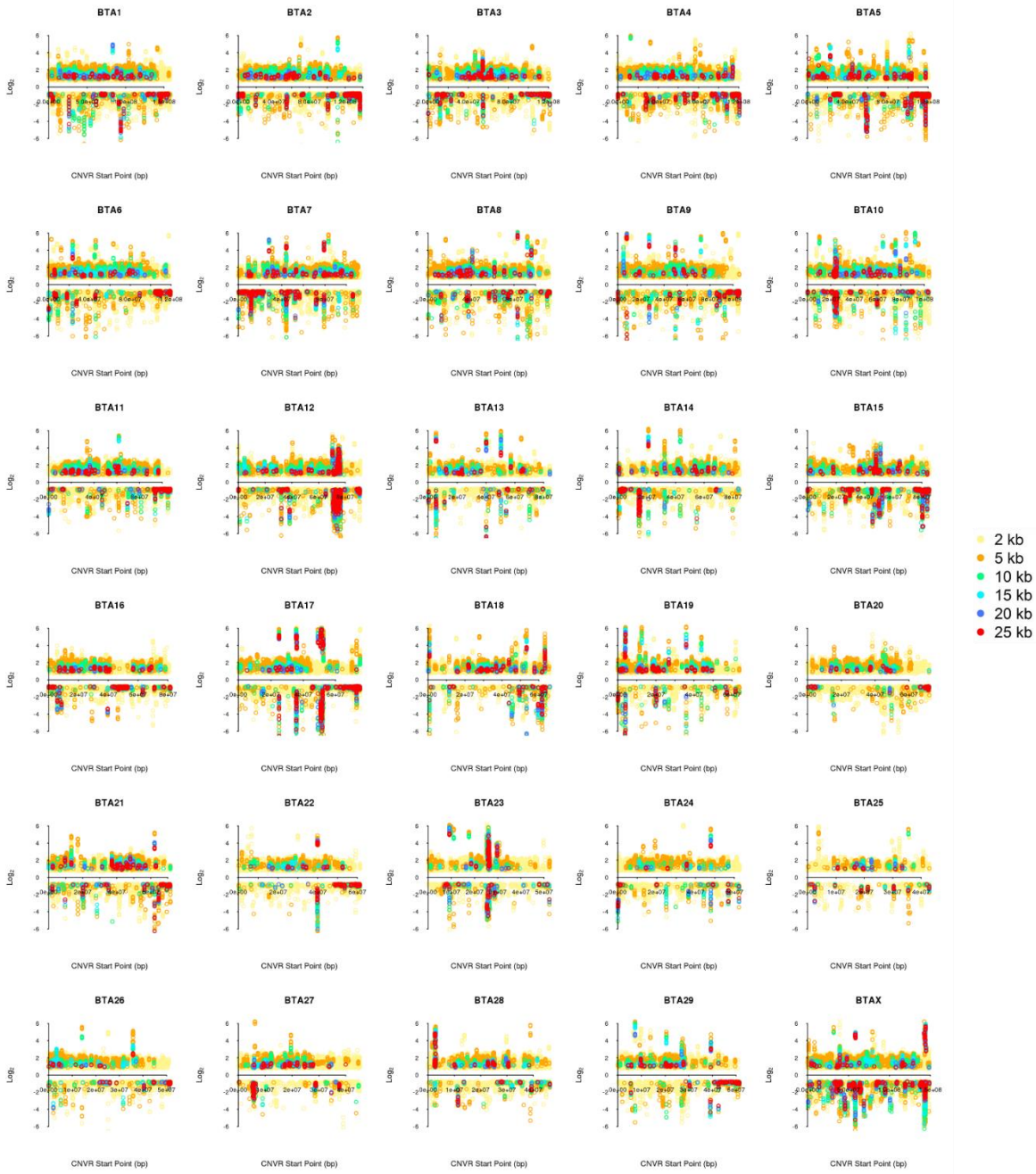


Figure A-3 Comparison of CNVR detected by chromosome for window sizes from 2 kb to 25 kb.

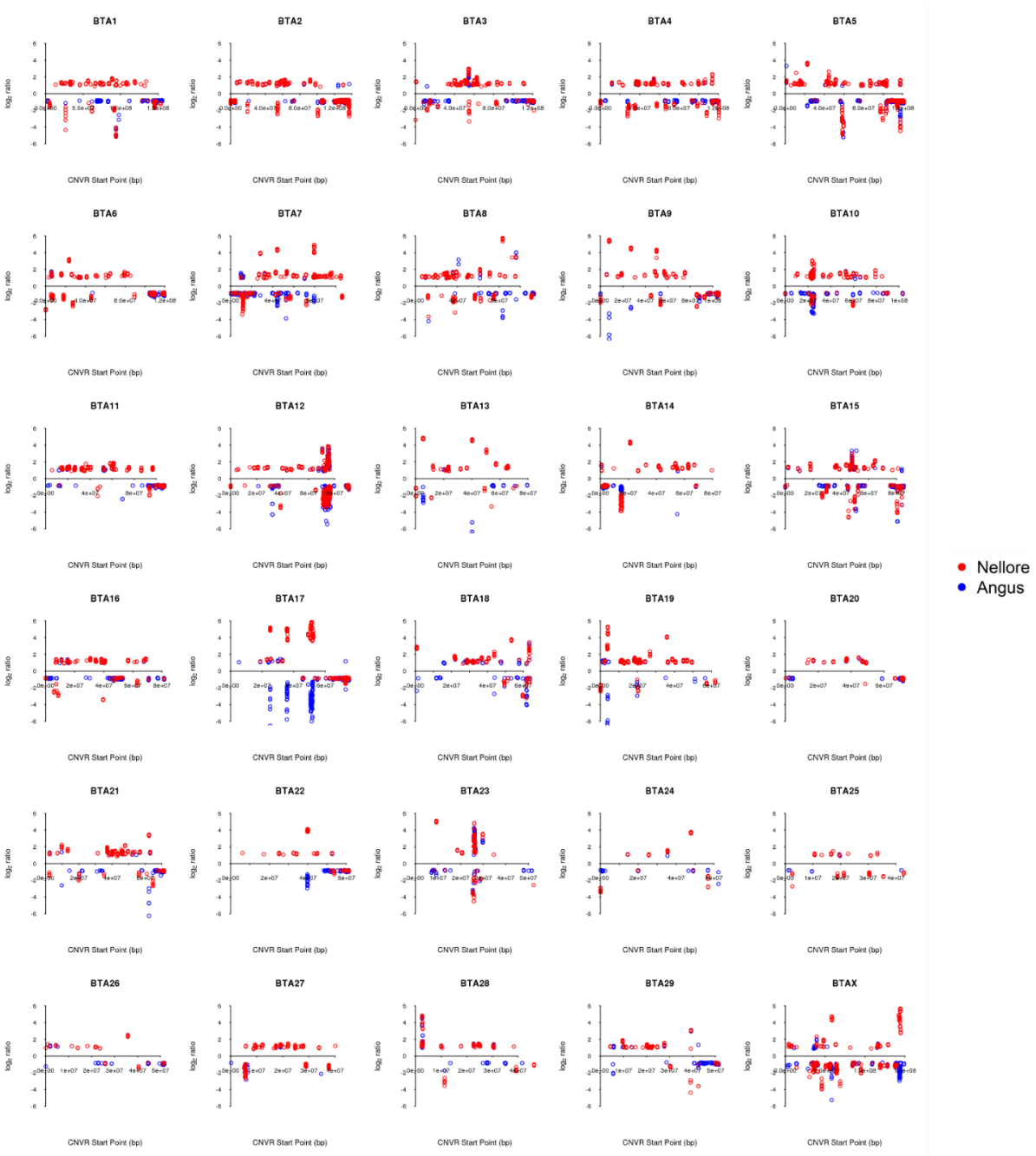


Figure A-4 CNV by chromosome identified in Nellore and Angus using a 25 kb minimum window.

Figure A-5 CNVR identified across the genome in all animals using different control animals.

Figure A-5 Continued.

Figure A-5 Continued.

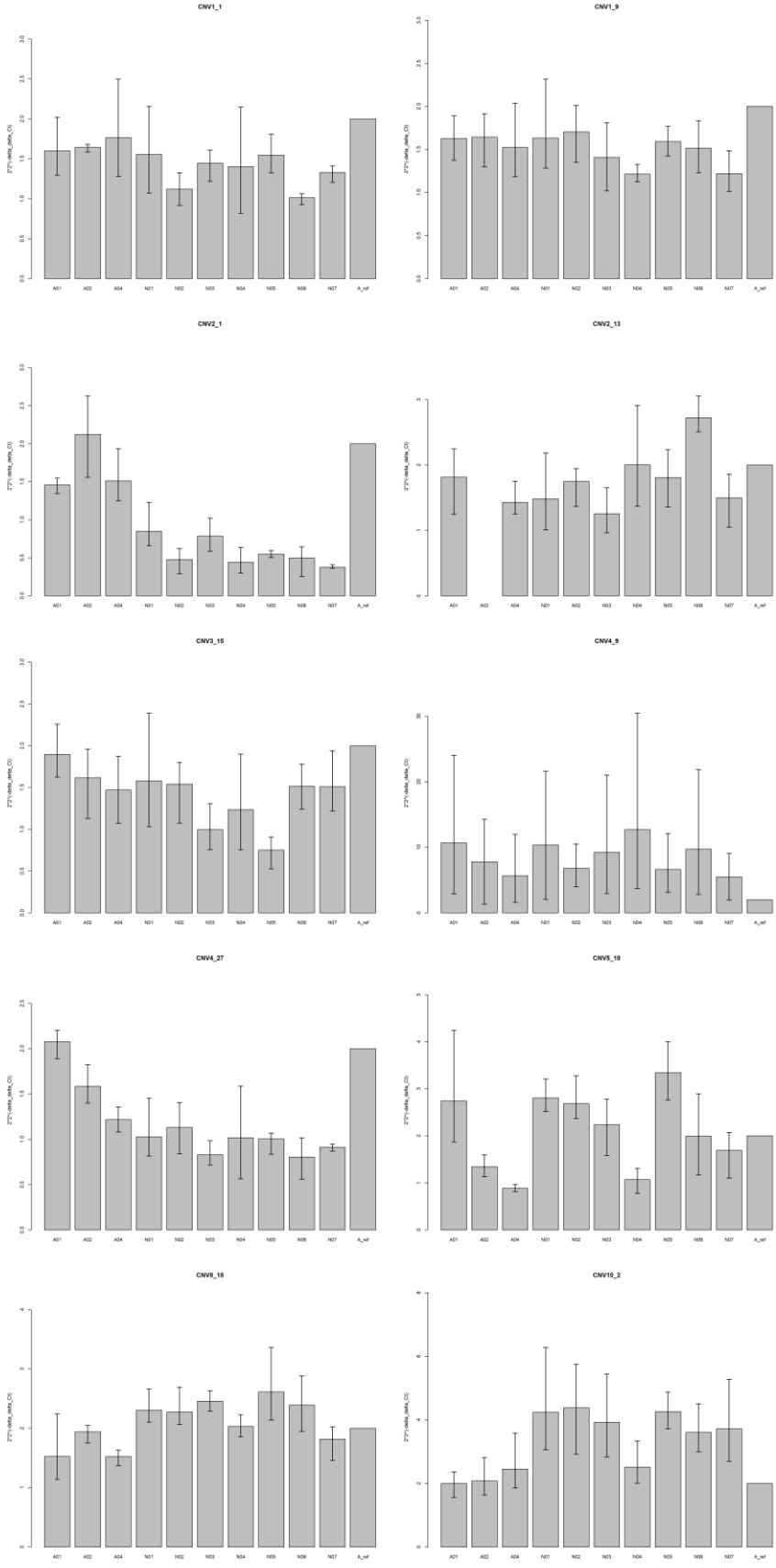


Figure A-6 Relative number of copies in each of the animals tested for each CNVR. In comparison to the control, a doubling of the number of copies in the test genome is equivalent to a log2 ratio of 1, half the number of copies is represented by log2 ratio of -1, and an unchanged number of copies is indicated by log2 ratio of 0.

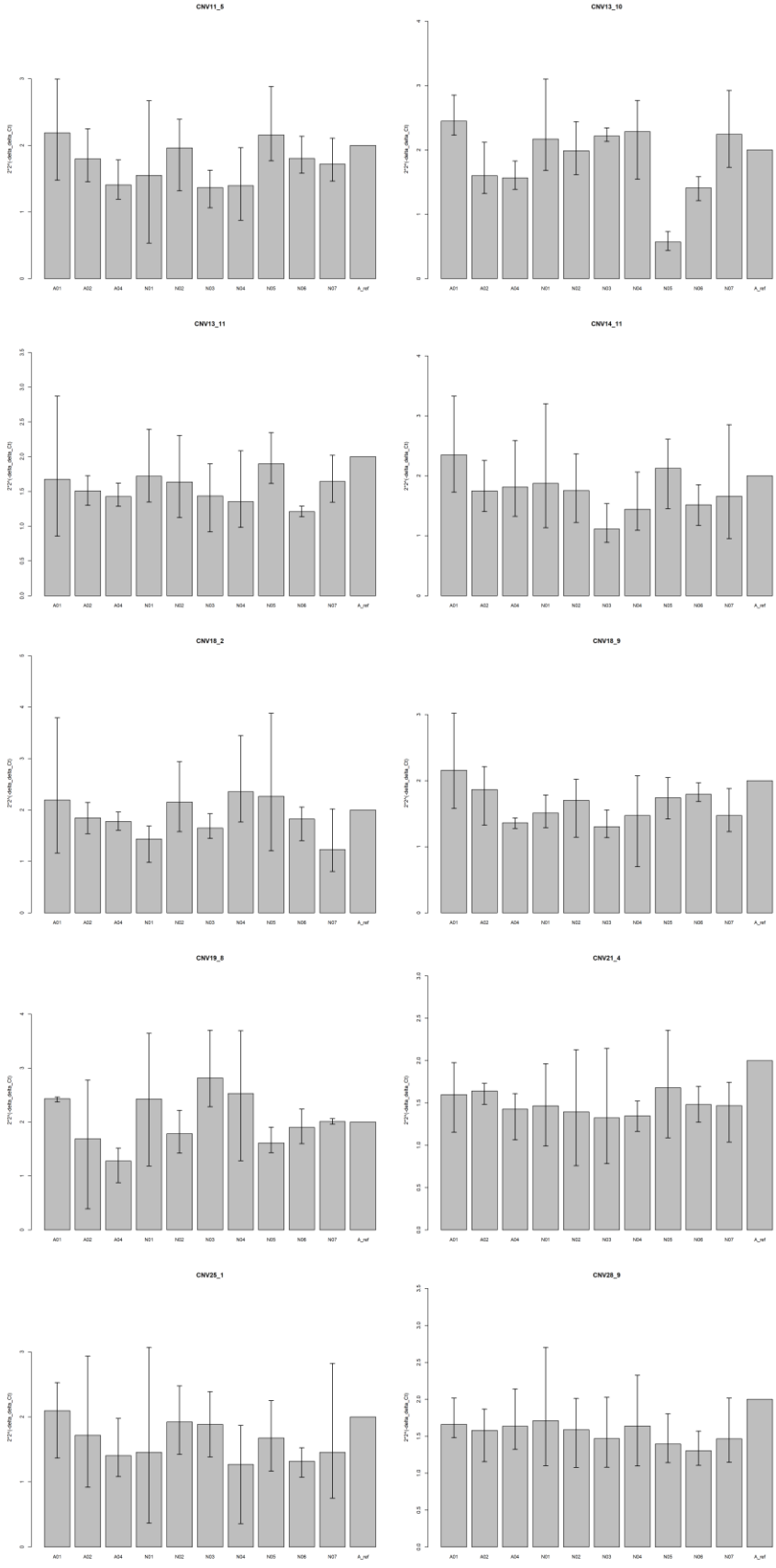


Figure A-6 Continued.

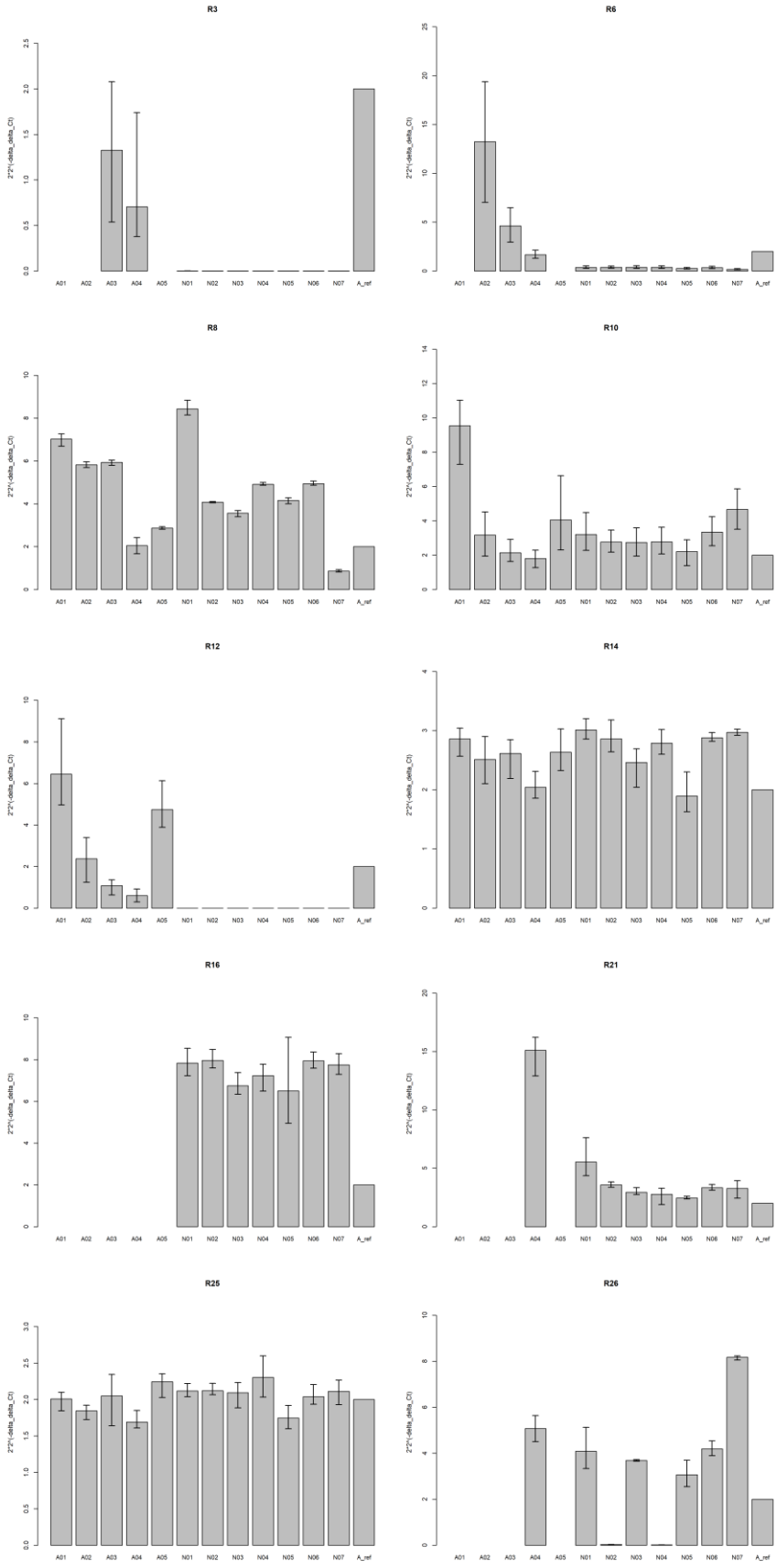


Figure A-7 Quantitative PCR validation for CNVR detected by the three control animals.

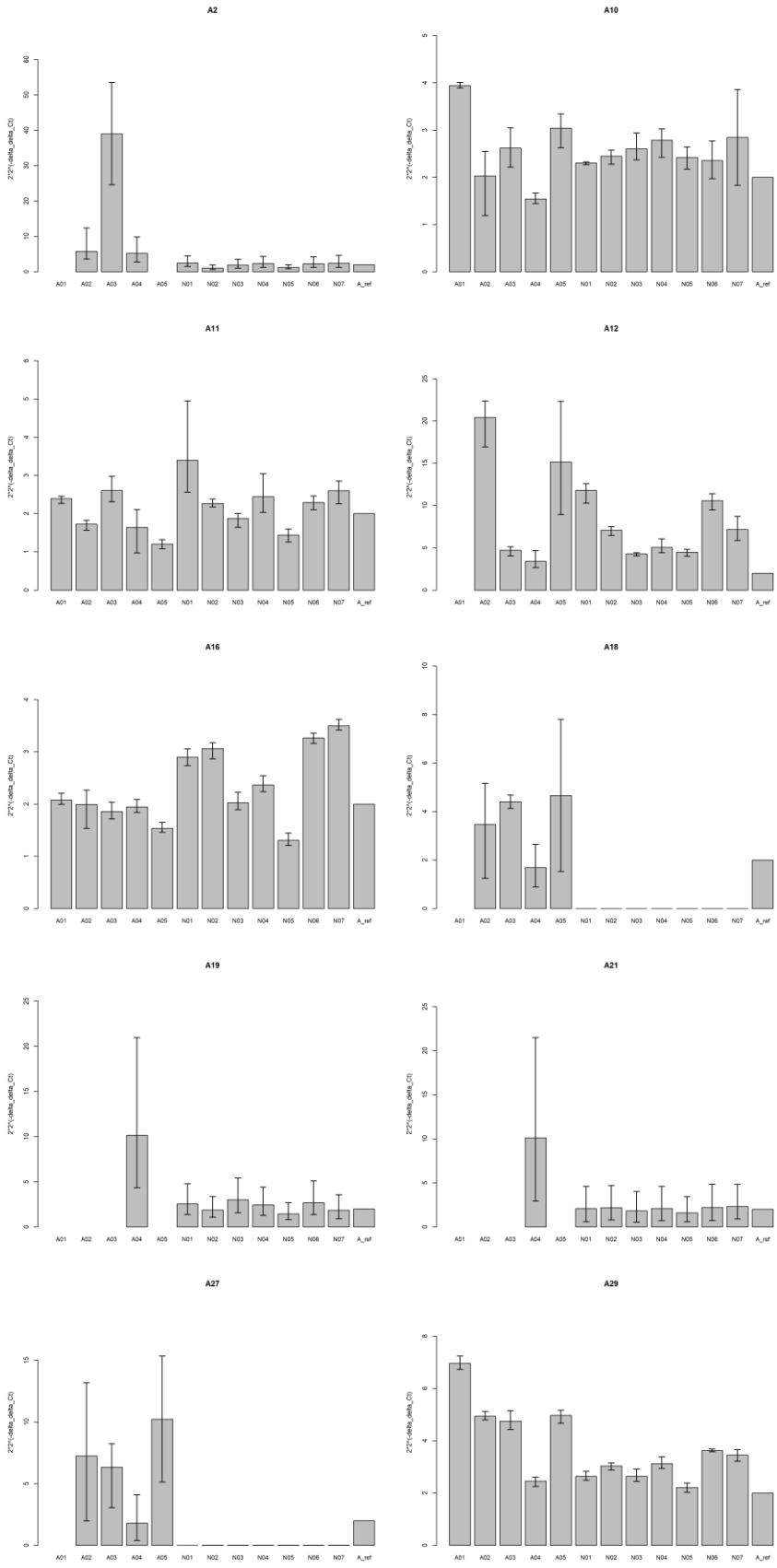


Figure A-8 Quantitative PCR validation for CNVR detected by the three software applications.

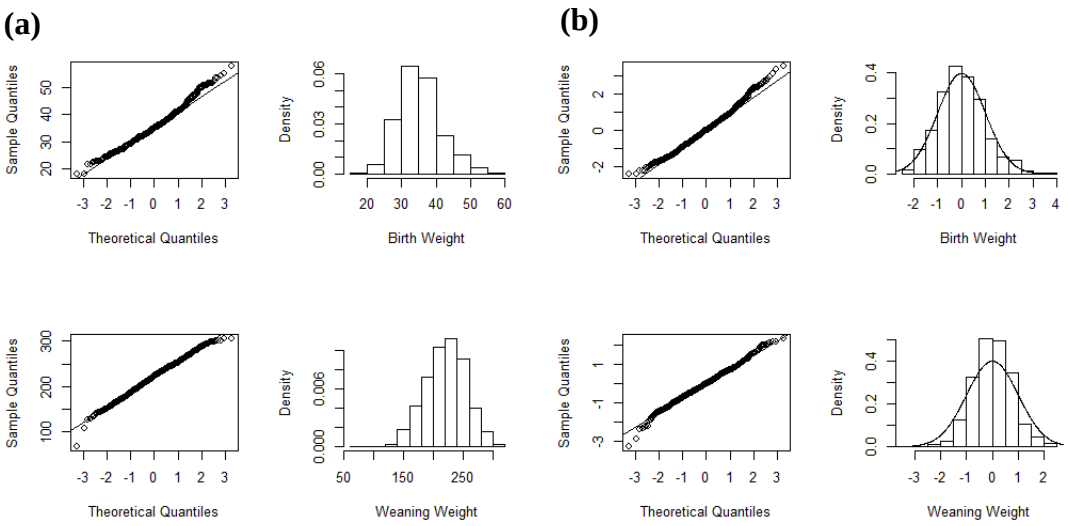


Figure A-9 Distribution of birth weight and weaning weight and their residuals. The Q-Q plots and histograms of a) birth weight and weaning weight, and b) their residuals from Proc Mixed in SAS are summarized. The first row shows birth weight and the second row shows weaning weight. The curves on histograms are standardized normal probability density functions.

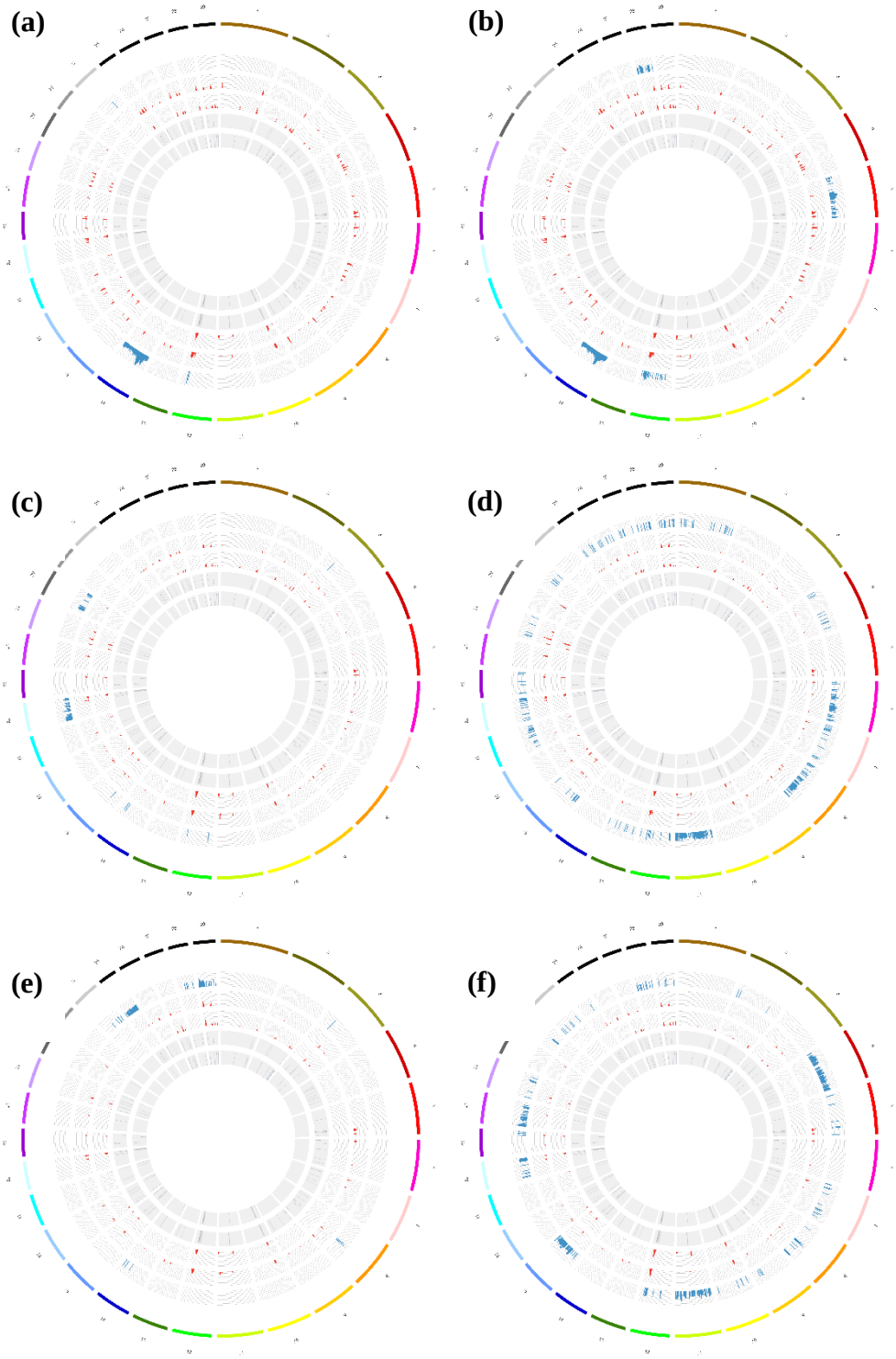


Figure A-10 Circus plots of the $-\log_{10} P-$ values before Benjamini-Hochberg correction.

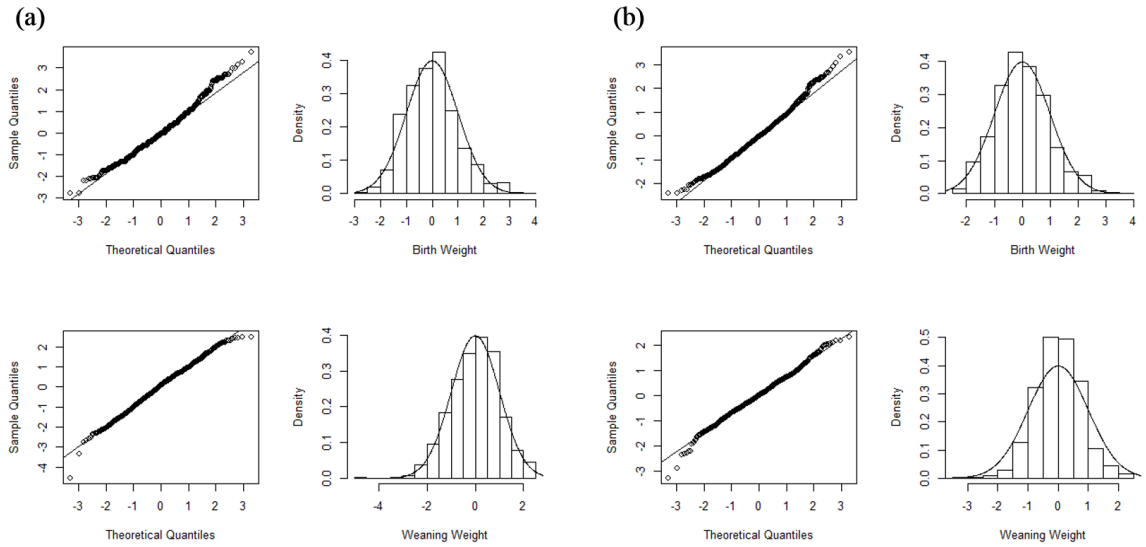


Figure A-11 Distribution of standardized birth weight and weaning weight, and their residuals. The Q-Q plots and histograms of a) standardized birth weight and weaning weight, and b) their residuals from a MLR model with sex, birth season and weaning age are summarized. The first row shows birth weight and the second row shows weaning weight. The curves on histograms are standardized normal probability density functions.

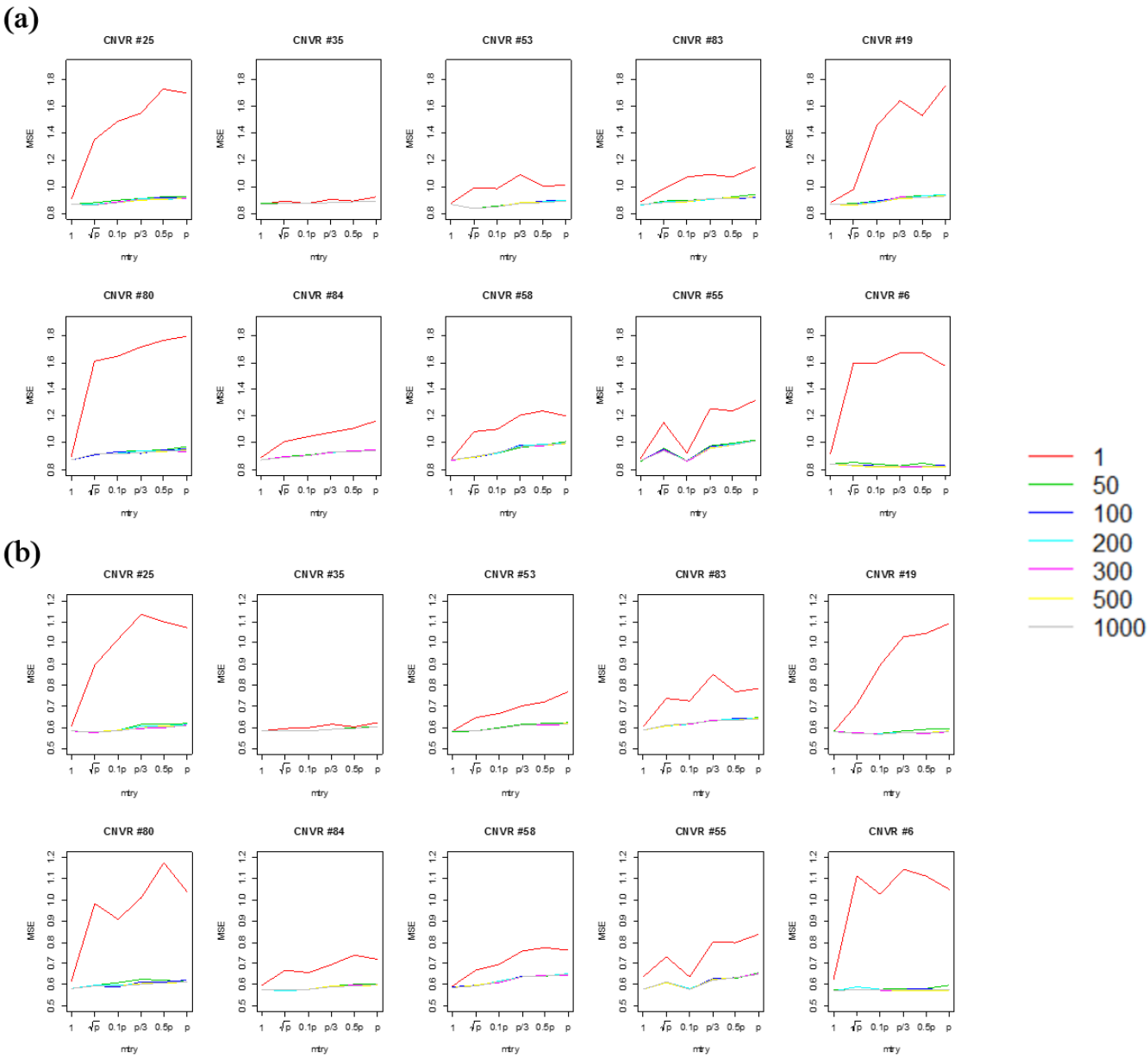


Figure A-12 MSEs of randomly selected CNVR for RF model from CV. The MSEs of a) birth weight and b) weaning weight for 10 randomly selected CNVR from CNV-SNP set 1 fitting the RF model for CV are plotted. Different colors indicate different number of trees to grow (ntree).

APPENDIX B

TABLES

Table B-1 CNVR counts and nucleotide content (Mb) by window size and animal.

Sample		2 kb		5 kb		5 kb*		10 kb		10 kb*		15 kb		20 kb		25 kb	
		Count	**Mb**	**Count**	**Mb**	**Count**	**Mb**	**Count**	**Mb**	**Count**	**Mb**	**Count**	**Mb**	**Count**	**Mb**	**Count**	**Mb**
Angus	A01	145286	502	21085	164	20618	163	6588	109	13442	232	3795	96	2584	87	1914	81
	A02	101990	323	11368	79	10253	74	2596	39	7001	106	1373	32	875	27	628	25
	A03	42781	134	4205	32	3737	30	746	14	871	18	337	11	233	10	165	9
	A04	4346	21	1032	12	1079	13	466	11	492	13	271	10	198	9	141	8
	A05	131496	433	16924	123	13240	110	4772	76	11065	179	2648	64	1734	57	1271	52
Nellore	N01	140007	469	18411	140	18842	141	5311	88	12623	210	3023	76	2037	69	1518	64
	N02	102482	333	11586	89	12137	91	2632	46	7403	118	1371	38	916	35	679	32
	N03	106661	343	11960	88	30603	148	2613	43	7916	120	1334	34	841	29	594	26
	N04	97932	317	10865	83	10367	81	2360	42	6617	105	1212	34	818	31	603	29
	N05	105259	341	11986	90	11678	88	2766	46	7910	124	1448	38	933	34	712	32
	N06	102149	332	11610	88	7676	71	2678	46	7473	118	1401	37	902	33	697	31
	N07	109715	353	12325	92	6583	67	2945	49	8257	130	1586	41	1056	37	790	34

2 kb: window-size = 1000, consecutive-window = 4; 5 kb: window-size = 1000, consecutive-window = 10; 5 kb*: bigger-window = 5, consecutive-window = 10; 10 kb: window-size = 2000, consecutive-window = 10; 10 kb*: window-size = 5000, consecutive-window = 4; 15 kb: window-size = 3000, consecutive-window = 10; 20 kb: window-size = 4000, consecutive-window = 10; 25 kb: window-size = 5000, consecutive-window = 10.

Table B-2 Depth of coverage of whole genome sequences and associated CNVR counts by animal.

Sample	Coverage	CNVR count
A01	45x	20618
A02	42x	10253
A03	43x	3737
A04	48x	1079
A05	39x	13240
N01	47x	18842
N02	47x	12137
N03	88x	30603
N04	45x	10367
N05	45x	11678
N06	36x	7676
N07	33x	6583
A_ref	75x	-

Table B-3 Number of CNVR per autosome in Nellore and Angus for different window sizes.

Chr	Nellore						Angus					
	2 kb	5 kb	10 kb	15 kb	20 kb	25 kb	2 kb	5 kb	10 kb	15 kb	20 kb	25 kb
1	56185	6611	1128	503	377	183	30433	3749	918	481	317	203
2	48997	5976	1424	741	533	409	25412	3639	991	558	410	307
3	36552	4591	1104	546	327	243	20856	2873	771	433	276	196
4	40436	4783	1025	566	340	266	22205	2899	791	473	282	207
5	36249	4569	1156	605	392	313	20723	3153	993	589	376	296
6	43096	5059	1076	619	355	264	21485	2655	640	381	236	182
7	35100	3996	1005	577	359	282	20323	2818	961	575	387	288
8	33801	4061	1027	514	318	242	17408	1896	459	210	145	96
9	33892	3412	724	339	215	177	17600	1909	428	223	145	103
10	29076	3440	887	542	318	208	16294	2005	580	319	189	132
11	28815	3632	959	444	314	245	17143	2405	700	345	247	200
12	30583	4282	1141	662	450	375	17328	2704	770	452	297	242
13	14538	1343	299	197	129	102	8466	769	198	121	69	53
14	22676	2620	602	261	185	134	12680	1585	449	246	175	123
15	24379	2771	719	390	288	214	14425	1807	526	321	222	160
16	21142	2276	540	273	196	147	12495	1609	484	269	200	147
17	22956	2742	632	364	230	172	13380	2200	715	409	300	231
18	12663	1675	455	268	202	143	7941	1059	309	186	131	99
19	11069	1426	472	337	208	190	6886	893	253	159	107	77
20	21053	1952	408	163	123	81	11641	1265	304	141	96	69
21	17864	2191	567	327	236	173	10676	1419	391	198	152	107
22	13110	1287	301	139	83	56	7957	951	303	182	119	79
23	10735	1067	281	150	103	83	6725	771	243	144	99	82
24	14641	1141	173	112	72	37	8612	758	139	74	50	33
25	4881	549	195	96	82	54	2674	234	67	28	25	12
26	13037	1231	273	153	44	34	7952	885	222	123	55	37
27	14766	1783	556	282	228	152	6059	586	139	60	53	29
28	12513	1518	380	225	134	110	6906	826	201	107	67	50
29	12759	1459	357	191	129	103	7477	899	321	165	108	89
X	46641	5297	1439	789	533	401	25737	3183	883	452	289	190
Sum	764205	88740	21305	11375	7503	5593	425899	54404	15149	8424	5624	4119

2 kb: window-size = 1000, consecutive-window = 4; 5 kb: window-size = 1000, consecutive-window = 10; 10 kb: window-size = 2000, consecutive-window = 4; 15 kb: window-size = 3000, consecutive-window = 10; 20 kb: window-size = 4000, consecutive-window = 10; 25 kb: window-size = 5000, consecutive-window = 10.

Table B-4 Proportion of CNVR separated by less than a step.

Sample	5 kb <1 kb	10 kb <2 kb	15 kb <3 kb	20 kb <4 kb	25 kb <5 kb
A01	0.043	0.053	0.054	0.049	0.044
A02	0.020	0.035	0.039	0.032	0.027
A03	0.025	0.060	0.080	0.060	0.098
A04	0.081	0.088	0.089	0.102	0.100
A05	0.037	0.046	0.051	0.037	0.047
N01	0.034	0.043	0.042	0.037	0.044
N02	0.021	0.034	0.038	0.033	0.034
N03	0.017	0.026	0.030	0.024	0.024
N04	0.019	0.035	0.037	0.037	0.030
N05	0.020	0.034	0.036	0.031	0.031
N06	0.020	0.031	0.031	0.036	0.033
N07	0.021	0.032	0.033	0.036	0.032

Table B-5 Comparison of length of CNVR identified by BD, CNV-seq and RS before filtration.

Sample	Number			Mean (kb)			Median (kb)		
	BD	**CNV-seq**	**RS**	**BD**	**CNV-seq**	**RS**	**BD**	**CNV-seq**	**RS**
A01	16488	1914	8733	1022.79	42.16	15.66	0.68	32.5	0.15
A02	18237	628	36938	1058.20	39.73	3.86	0.62	30	0.19
A03	17652	165	9517	1190.51	54	15.54	0.70	40	0.16
A04	16018	141	10360	1306.24	59.27	13.68	0.78	37.5	0.13
A05	18211	1271	8404	1230.63	41.24	14.68	0.67	32.5	0.15
N01	21521	1518	11153	968.84	42.24	11.51	0.65	32.5	0.18
N02	22017	679	40099	1008.25	46.97	4.32	0.65	30	0.19
N03	21092	594	43922	1044.59	44.57	4.30	0.73	30	0.20
N04	18546	603	41273	1038.73	47.64	4.41	0.70	32.5	0.19
N05	19297	712	21012	979.40	44.95	7.04	0.65	32.5	0.19
N06	18790	697	10892	1041.12	44.62	15.11	0.656	30	0.19
N07	22548	790	35481	938.35	43.45	4.68	0.62	32.5	0.19
A_ref	-	-	10741	-	-	12.46	-	-	0.13

Table B-6 SNP coding for additive, dominant and recessive models. Missing genotypes were coded as 0. There were 0.014% SNP with missing genotypes (1800~3600 out of 25412474).

	Additive model	Dominant model	Recessive model
A1A1	2 2	2 2	2 2
A2A2	1 1	1 1	1 1
A1A2	2 1	1 1	2 2
A2A1	1 2	1 1	2 2
Unphased heterozygous	1 2	1 1	2 2

Table B-7 Further SNP coding in additive, dominant and recessive models for MLR, RF and RT models.

	2 2	**1 2 or 2 1**	**1 1**
Additive model	2	1	0
Dominant model	1	1	0
Recessive model	1	0	0

Table B-8 CNVR collections having MSEs < baseline MSEs of both birth weight and weaning weight for MLR model.

CNVR-SNP set	CNVR collections having MSEs < baseline MSEs of both birth weight and weaning weight
1	1*2, 1*3, 3*3, 2*3, 2*4
2	2*2, 3*2, 6*2, 7*2, 8*2, 9*2, 10*2, 12*2, 13*2, 2*3, 3*3, 5*3, 6*3, 7*3, 8*3, 9*3, 10*3, 12*3, 13*3, 2*4, 3*4, 6*4, 7*4, 8*4, 9*4, 10*4, 12*4, 13*4, 6*5, 8*5, 7*13, 4*15, 5*16, 6*16, 8*16, 10*16, 6*17, 7*17, 4*18, 9*18
3	14*11
4	9*9, 14*9, 8*10, 13*12, 18*12, 3*13, 8*13, 13*13, 5*14, 15*14, 1*17, 6*17, 5*21, 10*14, 10*21, 15*21, 10*22, 15*22, 3*23, 8*23, 9*23, 11*23, 13*23, 14*23, 16*23, 20*23, 12*26
5	1*1
6	-

CNVR collections having MSEs < baseline MSEs of both birth weight and weaning weight for MLR model are summarized.

Table B-9 Prediction in testing data set using BSLMM, RT and RF models.

Phenotype	CNVR-SNP set	Model		
		BSLMM	**RT**	**RF**
Birth weight	1	0.99/0.27/1.45/20.97	1/0.24/-0.2/1.98	0.96/0.31/1.37/23.28
	2	0.98/0.27/1.37/23.52	1.07/0.15/-0.66/3.24	1.01/0.23/1.11/63.47
	3	0.99/0.27/1.22/35.48	1.03/0.21/5.41/8.7	1/0.25/1.89/14.2
	4	0.98/0.3/1.01/571.24	1.03/0.23/4.41/9.11	0.97/0.31/1.02/276.7
	5	0.95/0.31/1.19/39.03	1.02/0.22/2.36/11.89	1/0.25/1.11/66.37
	6	0.99/0.27/1.35/24.79	1.01/0.24/-0.02/1.03	1.02/0.21/-0.05/1.15
Weaning weight	1	0.52/0.68/1.24/31.26	0.57/0.64/-0.45/2.82	0.52/0.68/1.02/279.39
	2	0.55/0.65/1.55/18.35	0.58/0.64/-0.04/1.05	0.55/0.66/1.23/33.69
	3	0.56/0.65/1.34/25.31	0.56/0.65/-0.04/1.16	0.57/0.64/1.26/31.82
	4	0.52/0.68/1.01/565.76	0.54/0.67/-0.57/3.32	0.52/0.68/1.06/108.68
	5	0.56/0.65/1.58/17.91	0.58/0.64/-0.05/1.07	0.54/0.66/1.54/18.47
	6	0.55/0.65/1.4/22.71	0.59/0.63/-2.39/5.31	0.57/0.65/1.3/28.41

MSEs of prediction in testing data set for BSLMM model, regression tree model and random forest model for each of the CNVR-SNP set are summarized in this table. The minimum MSE in each category were labeled in bold.

Table B-10 Standard deviations of MSEs from bootstrap to access variability of model performance estimates.

CNVR-SNP set	Birth Weight				Weaning Weight			
	BSLMM	MLR	RT	RF	BSLMM	MLR	RT	RF
1	0.17	0.06 (0.01, 6)	0.06	0.01	0.12	0.04 (0.01, 6)	0.04	0.00
2	0.21	0.06 (0.01, 42)	0.06	0.00	0.11	0.04 (0.01, 42)	0.04	0.00
3	0.15	0.06 (0.01, 3)	0.06	0.01	0.13	0.05 (0.00, 3)	0.04	0.00
4	0.11	0.04 (0.00,28)	0.06	0.00	0.08	0.03 (0.00,28)	0.04	0.00
5	0.17	0.07 (0.01, 2)	0.06	0.01	0.13	0.05 (0.01, 2)	0.04	0.00
6	0.19	0.07 (0.01, 2)	0.06	0.05	0.13	0.05 (0.01, 2)	0.04	0.00

Standard deviations of MSEs (SD(MSE)) from bootstrap were summarized for BSLMM, MLR, RT and RF models to assess variability of model performance estimates. Note that the standard deviations of MSEs from BSLMM model were obtained using a different method compared to standard bootstrap used in other models. For MLR, the first number is the mean of SD(MSE) for all possible best CNVR collections, and the numbers in parentheses are the standard deviation of SD(MSE) and number of possible best CNVR collections.

APPENDIX C

ADDITIONAL FILES

Additional File C-1 Information and qPCR primers for CNVR validated.

Additional Files C-2-C9 Lists of CNVR identified by each of the minimum detectable CNVR sizes. A list of CNVR identified by chromosome for each of the animals for each of the minimum detectable CNVR sizes are reported. In each of the tables, the column headers are: CNVR name, chromosome, start, end, size, log2 ratio and p-value.

File name (.xlsx)	Minimum detectable CNV size	Parameters
Additional File C-2	2 kb	window-size = 1000 consecutive-window = 4
Additional File C-3	5 kb	window-size = 1000 consecutive-window = 10
Additional File C-4	5 kb	bigger-window = 5 consecutive-window = 10
Additional File C-5	10 kb	window-size = 2000 consecutive-window = 10
Additional File C-6	10 kb	window-size = 5000 consecutive-window = 4
Additional File C-7	15 kb	window-size = 3000 consecutive-window = 10
Additional File C-8	20 kb	window-size = 4000 consecutive-window = 10
Additional File C-9	25 kb	window-size = 5000 consecutive-window = 10

Additional Files C10 The CNVR lists using Angus animal (A03) as control animal. The minimum detectable CNVR size for them is 25 kb. In each of the tables, the column headers are: CNVR name, chromosome, start, end, size, log2 ratio and p-value.

Additional Files C11 The CNVR lists using Nellore animal (N05) as control animal. The minimum detectable CNVR size for them is 25 kb. In each of the tables, the column headers are: CNVR name, chromosome, start, end, size, log2 ratio and p-value.

Additional File C-12 Lists of CNVR mapped to RefSeq genes in Nellore and Angus animals. Lists of CNVR identified by all three controls that overlap with RefSeq genes are summarized. The RefSeq genes are downloaded from UCSC genome database. Only genes are directly overlap with CNVR are reported.

Additional File C-13 The enriched GO terms in Nellore and Angus animals. The enriched GO terms in Nellore and Angus animals are summarized. DAID was used for GO enrichment analysis for genes that overlap with CNVR identified by all three controls.

Additional File C-14 CNVR overlapping with QTL in Nellore and Angus animals. The CNVR overlapping with QTL from AnimalQTLdb are summarized for Nellore and Angus animals. CNVR were identified by all three controls.

Additional File C-15 Quantitative PCR results.

Additional File C-16 Common CNR set detected by the three control animals.

Additional File C-17 CNVR identified by BreakDancer and RAPTR-SV.

Additional File C-18 Common CNR set detected by the three software applications.

Additional File C-19 Phenotype data of 995 animals from McGregor Genomics beef cattle population.

Additional File C-20 The two sets of common CNVR and their RCN.

Additional File C-21 The six CNVR-SNP sets.

Additional File C-22 The IDs of CNVR associated with each SNP model and their location in the genome for each of the CNVR-SNP sets.

Additional File C-23 The P-values of the significant SNP after Benjamini-Hochberg correction for birth weight.

Additional File C-24 The P-values of the significant SNP after Benjamini-Hochberg correction for weaning weight.

Additional File C-25 The RefSeq genes overlapping with CNVR-tagged SNP.

Additional File C-26 Details of how the covariates were coded in the MLR model.

Additional File C-27 The CNVR used in BSLMM, MLR, RT and RF models.

Additional File C-28 The details of the RefSeq genes overlapping with best collections of CNVR for each model.

Additional File C-29 The top genes overlapping with most CNVR for each model.