SEQUENCING THE GENOME OF THE NORTH AMERICAN BISON

A Dissertation

by

LAUREN KRISTEN DOBSON

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,	James N. Derr
Committee Members,	James E. Womack
	Terje Raudsepp
	Scott N. Dindot
	David G. Riley
Interdisciplinary Faculty Chair,	Craig J. Coates

August 2015

Major Subject: Genetics

ABSTRACT

American bison *(Bison bison*) is a well-known iconic species with a history and legacy intertwined with the Plains of North America. Unfortunately, the American colonization of North America in the late 1800's resulted in the almost complete destruction of the American bison and subsequent population bottleneck. Bison were also faced with forced hybridization of domestic cattle genetics *(Bos taurus)*, through failed experiments of some ranchers to produce a hardier beef animal for the great-plains. The hybridization of domestic cattle into bison presents challenges in the management and conservation of American bison today, primarily because it is difficult to differentiate between hybrid cattle-bison and purebred bison within a population.

Whole genome sequencing provides the next step in advancing bison management and conservation. A 2.82-Gb *de novo* reference assembly of the American bison genome was produced using approximately 75X coverage, utilizing both mate pair and pair-end sequencing. Illumina, Inc. and 454 Life Sciences Technologies raw sequence reads were mapped to both nuclear and mitochondrial sequences of the domestic cattle reference UMD3.1 (Ensembl GCA_000003055.3), in order to detect genetic variants, including single nucleotide variants (SNVs) and insertion and deletions (INDELs). An additional 14 re-sequenced bison genomes were also aligned to the UMD3.1 domestic cattle reference sequence to identify genomic variants. These variants were determined and annotated to examine their effect on gene structure and function in bison.

With the completed *de novo* plains bison reference genome sequence a comparison of historic and modern bison sequences identified genomic variants and were compared across bison populations. Historic bison samples that predate cattle and bison introgression were sequenced and conserved genomic regions between historic and current bison were identified. Identified variants between modern and historic bison provided an outline of the genetic architecture of bison that existed before the population bottleneck. This genomic analysis of North American bison provides insight into the genetic history, taxonomy, and inheritance of important genetic traits in bison that have allowed them to not only survive but thrive in their recovery from this population bottle neck that occurred 130 years ago.

DEDICATION

I dedicate my dissertation work to all of the people who have supported and encouraged my research over the years. To the bison producers who have shown such passion for this animal, that it continues to inspire me to produce the best work that I can. I am especially grateful for my loving and patient parents, Parker and Sharon Dobson, who always pushed me to never give up and to strive to do the best that I could to achieve my dreams.

ACKNOWLEDGEMENTS

I owe a sincere appreciation and gratitude to my advisor, Dr. James N. Derr, for his guidance and assistance to see this dissertation through from inception to completion. Thank you for providing me the many opportunities to attend meetings to meet different people who play roles in the bison industry and research, and for allowing me to be a part of so many different research projects and the service lab, which increased my interaction with bison producers and herd managers.

I would like to thank my committee members, Dr. James Womack, Dr. Terje Raudsepp, Dr. Scott Dindot, and Dr. David Riley, for their guidance and support throughout the course of this research. Thank you to my collaborators as well, Dr. Aleksey Zimin (University of Maryland), Dr. James Reecy, Dr. David Alt, and Dr. Julie Blanchong (Iowa State University), Dr. Tim Smith (USDA-MARC), and Dr. Steve Olsen (USDA-NADC), for additional support and communications throughout this process.

I am very grateful to Mr. Ted Turner and Dr. Dave Hunter for letting us collect samples from Templeton on the Green Ranch in Montana. If samples could not be obtained from the proper animal, this research would not have been as thorough or complete, so thank you. Thanks goes to the institutions that also provide the funding and resources that was needed to complete this extensive project.

I utilized the Texas A&M Institute for Genome Sciences and Society (formerly the Whole Systems Genomics Institute) server and software for analyses with the help of

Mr. Kranti Konganti. I would not have been able to perform such high-performance computing of my extensive data without this resource or support for aiding me in this research.

I am grateful to my close friends for encouraging me during this process and giving me their help and support when the time came for it. Thank you to Floyd Barnes, lab mate, for letting me run ideas by you and being with me from the start of this program. I would also like to thank fellow Genetics Graduate Students, Dr. Amanda Hulse-Kemp, Rachel Jordan, Yvette Halley, and Kylee Veazey for providing guidance and criticism when I needed it. Without their constant expression of pride and confidence in my ability to achieve a PhD, I would have not seen it through.

I express my love and affection to my parents, Mr. and Mrs. Parker Dobson, for giving me the strength to finish this project through overwhelming times while motivating me to be the best that I could. To my sister, Kelly Risko, who gave me the encouragement and guidance through her own experience of post graduate work to keep me motivated to finish. To all of my dogs who have been with me through this, most recently, Oakley, Remi, Buzz and Scooby, for allowing me to sacrifice time with them to complete this journey. To my patient, understanding, encouraging, and motivating boyfriend, Shaun Davis, thank you for getting me through this and helping me stay calm and always having confidence in me.

I have meet many different bison producers, herd managers, researchers, colleagues, and compassionate individuals who have helped guide and encourage me through this journey. To everyone I say “Thank you for always believing in me!”

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES xi

CHAPTER I INTRODUCTION 1

History of Bison 1
Challenges of Bison over the Last 150 Years 2
Status of North American Bison 6
Current Studies for North American Bison 9
Whole Genome Sequencing as a Solution 12
Study Objectives 15

CHAPTER II *DE NOVO* ASSEMBLY AND ANNOTATION OF THE NORTH AMERICAN BISON (*BISON BISON*) REFERENCE GENOME 16

Introduction 16
Materials and Methods 18
Results 28
Discussion 49

CHAPTER III 4 WAY GENOMICS COMPARISON OF NORTH AMERICAN BISON AND DOMESTIC CATTLE 52

Introduction 52
Materials and Methods 60
Results 66
Conclusions 144

CHAPTER IV CONCLUSIONS AND FUTURE RESEARCH 149

REFERENCES 155

APPENDIX 161

LIST OF FIGURES

Page

Figure 1. The bison reference genome animal Templeton.....19

Figure 2. Genomic contribution (percentage) of 8 core U.S. federal bison herds to Templeton.....28

Figure 3. Templeton's G-banded karyotypes showing normal diploid chromosome number 2n=60.....30

Figure 4. Identification to Templeton's X and Y chromosomes by FISH as marked by red fluorescence.....31

Figure 5. Variant (SNVs and INDELs) counts found for each chromosome from Templeton aligned to domestic cattle.....40

Figure 6. Variant counts with corresponding quality scores of SNVs (top) and INDELs (bottom) to evaluate the quality of variants annotated in Templeton.....41

Figure 7. Bison synteny to domestic cattle UMD3.1.76. Black anchors are those scaffolds that were found to have synteny with domestic cattle.....47

Figure 8. Bison scaffolds anchored to chromosome 1 of domestic cattle UMD3.1.76.....48

Figure 9. Agilent Tape Station results for assessing quality of genomic DNA of historical samples S6 and S9.....67

Figure 10. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for wood buffalo aligned to the reference bison.....73

Figure 11. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for CCSP bison aligned to Templeton.....77

Figure 12. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for YNP bison aligned to Templeton.....82

Figure 13. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for historic bison sample S6.....86

Figure 14. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for historic bison sample S9 87

Figure 15. Number of SNPs (top) and INDELs (bottom) by chromosome for each population when using Templeton's pseudo-chromosomes 94

Figure 16. Number of SNPs (blue) and INDELs (red) by chromosome for EIW samples mapped to domestic cattle UMD3.1 109

Figure 17. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for EIW buffalo 110

Figure 18. Number of SNPs (blue) and INDELs (red) by chromosome for CCSP bison mapped to domestic cattle UMD3.1 115

Figure 19. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for CCSP bison 116

Figure 20. Number of SNPs (blue) and INDELs (red) by chromosome for YNP bison mapped to domestic cattle UMD3.1 122

Figure 21. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for YNP bison 123

Figure 22. Number of SNVs (blue) and INDELs (red) by chromosome for historic bison samples mapped to domestic cattle UMD3.1 129

Figure 23. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) for historic bison samples 130

Figure 24. Phylogenetic tree from the combined VCF file for all 15 bison samples to UMD3.1 142

LIST OF TABLES

Page

Table 1. Statistics for Illumina (paired-end and mate pair) and 454 paired-end libraries used for *de novo* bison reference sequence 23

Table 2. List of loci used and genotypes for Templeton 29

Table 3. Global statistics (in base pairs) for Bison_UMD1.0 (NCBI) 32

Table 4. Gene and feature statistics (NCBI) 33

Table 5. Bison (UMD1.0) reference genome annotation comparison to domestic cattle (UMD3.1) and human (HuRef_1 and HuRef2 (GRCh38)) reference genome annotations 33

Table 6. Samtools flagstat statistics of the bison reference sequence mapped to UMD3.1 domestic cattle reference sequence 35

Table 7. Summary statistics for SNVs and INDELs found in Templeton compared to domestic cattle 37

Table 8. Base changes (SNVs) between Templeton and Domestic cattle 37

Table 9. Number of consequences (effect type) in genome of bison found after SNVs annotated between bison and domestic cattle 42

Table 10. Number of consequences (effect type) in genome of bison found after INDELs annotated between bison and domestic cattle 43

Table 11. Genomic regions associated with annotated SNVs and INDELs found in bison 44

Table 12. Biological functions of genes associated with annotated SNVs and INDELs in bison 44

Table 13. Chromosome summary from SyMap with number of bison scaffolds placed on each domestic cattle chromosome 46

Table 14. Sample information for the 15 bison samples used for sequencing analysis 66

Table 15. Partial blast report for historical sample S6 68

Table 16. Samtools flagstat statistics of the 4 EIW bison raw sequences mapped to the bison reference sequence UMD1.0 69

Table 17. Individual variant summary statistics of 4 wood buffalo found from comparing sequences to Templeton 70

Table 18. Base substitution counts for SNVs identified for 4 wood buffalo from alignment to Templeton 71

Table 19. Transition and transversion counts and ratio for each EIW sample and population total found against Templeton 72

Table 20. Number of common variants found between the 4 EIW samples 72

Table 21. Samtools flagstat statistics of the 4 CCSP bison raw sequences mapped to Templeton 74

Table 22. Individual variant summary statistics of 4 CCSP found from comparing sequences to Templeton 75

Table 23. Base substitution counts for 4 CCSP bison sequences aligned to Templeton 76

Table 24. Transition and transversion counts and ratio for each CCSP sample and population total found against Templeton 76

Table 25. Number of common variants found between the CCSP bison 77

Table 26. Samtools flagstat statistics of the 4 YNP bison raw sequences mapped to Templeton 79

Table 27. Individual variant summary statistics of 4 YNP bison found from comparing sequences to Templeton 79

Table 28. Base substitution counts for 4 YNP bison from alignment to Templeton 80

Table 29. Transition and transversion counts and ratio for each YNP sample and population total found against Templeton 81

Table 30. Number of common variants found between the 4YNP samples for SNPs and only 3 YNP samples for INDELs 82

Table 31. Samtools flagstat statistics of the 2 historic bison raw sequences mapped to Templeton 83

Table 32. Summary statistics of historical bison samples variants detected from alignment to Templeton 85

Table 33. Base substitution counts for historical bison from alignment to Templeton ... 85

Table 34. Transition and transversion counts and ratio for each historic bison sample found against Templeton. 85

Table 35. Summary of detected variants after alignment to Templeton by population ... 88

Table 36. Number of common variants to Templeton found between all individuals by populations 89

Table 37. Ratios of variants identified for all 4 populations to Templeton 89

Table 38. Comparison of variants identified for each population when analyzed from alignment to Templeton's scaffolds (S) or pseudo-chromosomes (C) 93

Table 39. Biological functions of genes associated with annotated SNPs for all 14 bison based on populations, or individuals for S6 and S9 93

Table 40. Unique genes after comparison of all bison genes found from annotation of variants identified to domestic cattle 95

Table 41. Samtools flagstat statistics of the 4 EIW bison raw sequences mapped to the domestic cattle reference sequence UMD3.1. 103

Table 42. Individual variant summary statistics of 4 EIW buffalo found from comparing sequences to domestic cattle UMD3.1 105

Table 43. Summary statistics for SNPs and INDELs found in 4 EIW buffalo compared to domestic cattle UMD3.1 106

Table 44. Number of common variants found between the 4 EIW samples 107

Table 45. Base changes (SNPs) between EIW buffalo and domestic cattle UMD3.1 ... 107

Table 46. Transition and transversion counts and ratio for each EIW sample and population totals found against domestic cattle reference UMD3.1 107

Table 47. Genomic regions associated with annotated SNPs and INDELs for combined EIW samples 108

Table 48. Biological functions of genes associated with annotated SNPs and INDELs in EIW buffalo 108

Table 49. Samtools flagstat statistics of the 4 CCSP bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 111

Table 50. Individual variant summary statistics of 4 CCSP found from comparing sequences to domestic cattle UMD3.1 113

Table 51. Summary statistics for SNPs and INDELs found in 4 CCSP bison compared to domestic cattle 113

Table 52. Number of common variants found between the 4 CCSP samples 114

Table 53. Base changes (SNPs) between CCSP bison and domestic cattle UMD3.1 114

Table 54. Transition and transversion counts and ratio for each CCSP sample and population totals found against domestic cattle reference UMD3.1 114

Table 55. Genomic regions associated with annotated SNPs and INDELs for combined CCSP samples 117

Table 56. Biological functions of genes associated with annotated SNPs and INDELs in CCSP 117

Table 57. Samtools flagstat statistics of the 4 YNP bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 118

Table 58. Individual variant summary statistics of 4 YNP found from comparing sequences to domestic cattle UMD3.1 119

Table 59. Summary statistics for SNPs and INDELs found in 4 YNP bison compared to domestic cattle 120

Table 60. Number of common variants found between the 4 YNP samples 121

Table 61. Base changes (SNPs) between YNP bison and domestic cattle UMD3.1 121

Table 62. Transition and transversion counts and ratio for each YNP sample and population totals found against domestic cattle reference UMD3.1 121

Table 63. Genomic regions associated with annotated SNPs and INDELs for combined YNP samples 124

Table 64. Biological functions of genes associated with annotated SNPs and INDELs in YNP 124

Table 65. Samtools flagstat statistics of the historic bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 126

Table 66. Individual summary statistics of 2 historic bison variants found from comparing sequences to domestic cattle UMD3.1 126

Table 67. Summary statistics for SNPs and INDELs found in 2 historical bison compared to domestic cattle UMD3.1 127

Table 68. Base substitutions (SNPs) and transition and transversions found historic bison samples and domestic cattle UMD3.1 128

Table 69. Genomic region associated with annotated SNPs and INDELs for both historical bison, S6 and S9 131

Table 70. Biological functions of genes associated with annotated SNPs and INDELs in 2 historical samples, S6 and S9 132

Table 71. Comparison across average amounts of each populations or samples SNPs detected to domestic cattle reference UMD3.1 134

Table 72. Comparison of those homozygous SNPs found to be in common for historic samples and how many of those are in common to certain populations 134

Table 73. Variant Summary for EIW, CCSP, and YNP populations, as well as the individuals Templeton, S6 and S9 136

Table 74. Unique annotated genes for YNP (Y), Templeton (T), and historic samples (O) after comparison of all bison gene lists found from annotation of variants identified to domestic cattle 138

Table 75. Unique genes after comparison of all bison genes found from annotation of variants identified to domestic cattle 139

Table 76. 12 significant GO terms resultant from gene list of those genes not annotated in S9 and then analyzed in DAVID for a GO analysis with the biological (GO) term, p-value and false discovery rater (FDR) 140

Table 77. Total variants found for bison populations and individuals to both UMD3.1 and UMD1.0 145

Table 78. Averages of SNPs idenfitifed for all of the 15 bison samples to UMD3.1 and UMD1.0 145

CHAPTER I

INTRODUCTION

History of Bison

Bison Classification

The *Bison* genus is represented by two extant species, *Bison bison* (North American bison) and *Bison bonasus* (European bison; Ward 2000). North American bison are a well-known iconic species that symbolize the early colonization of western North America. Their historical range is believed to have spanned about one-third of the entire content of North America, which extended north up into Canada and south down into Mexico (Hornaday 1886; Sanderson *et al.* 2008). Before colonization and importation of European domestic cattle breeds (*Bos taurus*) bison were known to be the most dominant and large herbivore in North America (McDonald 1981).

From a taxonomic point of view, North American bison are subdivided into two sub-species based on physical appearance and coat characteristics, wood buffalo (*Bison bison athabascae*) and plains bison (*Bison bison bison*; Hall, 1981; McDonald 1981; Meagher 1986). Wood buffalo is the common name and not wood bison, despite the fact that they are more closely related to bison than actual buffalo of Asian and African descent. Plains bison ranged historically across much of the United States and southwestern Canada, while wood buffalo occurred in north-western Canada. However the ranges of plains bison and wood buffalo most likely overlapped at times (Potter *et al.* 2010). Subspecies status was primarily assigned based on morphology (such as skulls,

horns, and body proportions, size and hair patterns), but there is not a consensus as to if these designations are valid; as previous genetic studies have not supported the distinction between plains and wood bison (McDonald 1981; Geist 1991; Cronin *et al*. 2013). European and North American bison also have different morphologies, but are more distinguishable than the differences between wood buffalo and plains bison.

American plains bison were once classified into two subspecies northern plains bison (*Bison bison montanae*) and southern plains bison (*Bison bison bison*) based on physical appearance, horn, and coat characteristics depicted in pre-1900s illustrations (Krumbiegel and Sehm 1989). Hornaday (1886) also noticed similar differences in coat characteristics and attributed this to geographical and climate influences. Even if these subspecies classifications were valid, since the 1900s they have crossbred freely and possibly eliminated these regional phenotypic differences (Coder 1975; Dary 1989; McHugh 1972). The only modern remnant of the southern plains bison is believed to be surviving animals in the Texas State Bison Herd, that were once reproductively isolated since its foundation by Charles Goodnight in the late 1800s until outside bulls were brought in to help increase genetic diversity in the herd in 2005 (Halbert 2003; Halbert 2004).

Challenges of Bison over the Last 150 Years

Population Bottlenecks in the Early 20th Century

North American bison have survived a number of historical population bottlenecks (Pertoldi *et al.* 2010). Based on personal observations, livestock census, carrying capacity calculations, and bison kill numbers, it is thought that approximately

30 to 60 million bison once populated North America (Seton 1937; Flores 1991; McHugh 1972; Roe 1970). Unfortunately, the American colonization of North America in the late 1800s resulted in the almost complete elimination of the American bison (both wood buffalo and plains bison) and lead to the subsequent population bottleneck, reducing the population size by over 99.9% in less than 100 years (Coder 1975; Dary 1989). Wood buffalo numbers have been down as low as 300 animals, relatives of the surviving animals are now found in the area belonging to Wood Buffalo National Park (Banfield and Novakowski 1960). Estimations of remaining plains bison ranged between a minimum of a few hundred individuals found in only 6 captive populations, to 500 to 600 left in the wild in the late 1800s (Hornaday 1913; Coder 1975; Halbert 2003). At this time, with both sub-species of North American bison in decline it was evident that extinction of the species was imminent and recovery efforts were needed (Halbert 2003).

Introgression with Domestic Cattle

The general consensus of the *Bison-Bos* genera split is that they once represented a single monophyletic clade believed to have derived from a common ancestor 0.5 to 2 million years ago in Eurasia (McDonald 1981). There is some disagreement over phylogenetic relationship among cattle and bison but most agree that the *Bison* genus should be included in the *Bos* genus (Simpson 1961; van Gelder 1977). Both extant bison species can produce viable offspring with not only domestic cattle (*Bos taurus*) but other members of the *Bos* genus (Meagher 1986). Female progeny are fertile, while loss of fertility of male hybrid offspring can be restored with repeated back-crossings (Verkaar *et al.* 2003).

It was well-known that the 5 cattlemen who helped with the recovery of bison also had bison for the purpose of creating hybrids with domestic cattle to produce a better meat source (Coder 1975). Hybridization, whether forced or spontaneous between bison and domestic cattle as well as other bovine species, may create animals with unique properties but at the same time compromise their genetic integrity (Verkaar *et al.* 2003). Hybridization of domestic cattle with bison presents challenges in the management and conservation of the American bison today, because most advanced generation backcrosses are morphologically indistinguishable from purebred bison (Douglas *et al.* 2011).

Genetic technologies and current research have confirmed this introgression of domestic cattle genetics, in both the mitochondrial (Polziehn *et al.* 1995; Ward *et al.* 1999) and nuclear DNA (Ward 2000; Halbert 2003; Halbert *et al.* 2005) in most modern bison herds. Freese *et al*. (2007) reported that at best less than 1.5% of the 500,000 plains bison in existence today can be considered as likely free of domestic cattle introgression. The effects of cattle introgression on bison physiology, behavior and fitness have been studied recently, but are still not fully understood (Freese *et al*. 2007).

Douglas *et al.* (2011) examined the complete mitochondrial DNA of 43 North American bison (samples with both bison and domestic cattle mitochondrial DNA) and compared it to 3 domestic cattle mitochondrial DNA sequences. Through this comparison, 642 fixed synonymous and 86 fixed non-synonymous differences were identified between bison and domestic cattle mitochondrial DNA out of 16,325 total nucleotides (Douglas *et al*. 2011). This finding validated that *Bos* and *Bison* species

have diverged about 1 Mya or less (Janecek *et al*. 1996; Buntjer *et al*. 2002). From a metabolic standpoint, bison and cattle exhibit differences, for example, in the winter bison are able to greatly reduce their metabolic rate, whereas cattle cannot do as efficiently (Freese *et al*. 2007). These non-synonymous mutations found in the mitochondrial DNA of bison-domestic cattle hybrids will most likely affect the mitochondrial function and overall fitness of the hybrids compared to non-hybrid bison (Douglas *et al*. 2011); therefore, those bison with domestic cattle mitochondrial DNA could be at a disadvantage if the cattle mitochondrial DNA affect bison energetics, growth and seasonal foraging behavior (Freese *et al*. 2007).

To assess if hybrid bison are at a fitness disadvantage, Derr *et al.* (2012) examined the weight and height differences between bison with bison mitochondrial DNA and domestic cattle mitochondrial DNA of 2 different bison populations, feedlot bison from Montana (nutritionally rich environment) and Santa Catalina Island (California; nutritionally stressful environment). In both environments it was shown that bison with domestic cattle mitochondrial DNA were on average smaller than bison with bison mitochondrial DNA (Derr *et al.* 2012). The association with domestic cattle mitochondrial DNA and reduced body size in bison was able to show the effects of genetic introgression from a different species can affect the phenotype of a hybrid species (Derr *et al.* 2012).

Status of North American Bison

Recovery of the North American Bison

North American bison population numbers started to increase due in part to recovery efforts of a small number of people in the 1880s (Halbert 2003). These individuals took it upon themselves to help save North American bison from extinction by capturing a few of the remaining wild North American bison and raising them in captivity (Ward 2000). There are 5 recognized populations, started by private ranchers that are responsible for playing a major role in the recovery of the North American bison. The foundation herds include the McKay-Alloway herd (1874) from Manitoba, Canada, Pablo-Allard (1873) herd located in Montana, the Dupree-Philip herd (1881) from South Dakota, the Charles Goodnight herd (1878) in Texas, and the Charles "Buffalo" Jones herd (1885) made up of individuals from Kansas, Nebraska, and Texas (Haley 1949; Ward 2000; Halbert 2003). There were also an additional 22 animals thought to have survived as wild bison in a 1902 census of remote areas of Yellowstone National Park (Garretson 1938; Meagher 1973; Coder 1975).

The National Zoological Park in Washington D.C. also aided in the recovery of North American bison with a mixture of bison from different herds and locations (Coder 1975). The collection of bison from 1888 to 1904 and the establishment of the National Zoological Park were overseen by Dr. William T. Hornaday, who also realized the importance of saving the North American bison from extinction (Coder 1975; Halbert 2003). Hornaday also collected bison hide, skull and skeletons that were archived at the

Smithsonian Institute's Natural Museum of Natural History in order to preserve the legacy of bison for future generations (Hornaday 1886).

Since the early 1900s, both the U.S. and Canadian governments have helped aid in the North American bison recovery by protecting the wild populations in Yellowstone and Wood Buffalo National Park from poachers (Ward 2000; Halbert 2003). With protection, the numbers of North American bison doubled from 1888 and 1902, and were considered safe from extinction in 1909 (Coder 1975). Bison numbers continued to increase rapidly from just above 2,000 in 1910 to over 21,000 bison in 1933 (Hornaday 1913; Seton 1937; Garretson 1938; Coder 1975).

Wood Buffalo National Park was estimated to contain 1,500-2,000 bison by 1922, but despite this steady population increase and with many objections from Canadian scientists, approximately 6,600 plains bison were moved into the herd from 1925-1928 (Banfield and Novakowski 1960; Roe 1970). This lead to the mixed breeding of wood buffalo and plains bison at Wood Buffalo National Park (van Camp 1989; Geist 1991), making it difficult to distinguish these hybrids from non-hybridized wood buffalo and plains bison. However, a pure sub-population of wood buffalo (Banfield and Novakowski 1960) was believed to have been used to establish populations at Mackenzie Bison Sanctuary and Elk Island National Park in Canada (Geist 1991).

Yellowstone National Park, founded in 1872, was the world's first National Park (Halbert 2003). However, poaching in Yellowstone National Park was widespread and President Cleveland enacted the Act to Protect the Birds and Animals in Yellowstone

National Park and to Punish Crimes in Said Park and For Other Purposes, to punish those that committed wildlife related crimes in the park (Dilsayer 1994; Freese *et al.* 2007). By 1902 there were only 22 remaining wild bison in Yellowstone National Park (Garretson 1938; Meagher 1973; Halbert 2003). In that year, President Roosevelt appointed Charles "Buffalo" Jones game warden to help preserve the wild bison in Yellowstone National Park. He played an integral part in supplementing additional outside animals from the Pablo-Allard (18 cows) and Charles Goodnight (3 bulls, but one died) herds into the Yellowstone herd (Garretson 1938; Coder 1975; Halbert 2003). The imported bison were confined to paddocks and were managed as a captive herd and once numbers increased in 1915 they were released into the park and able to interact with the "wild" bison (Meagher 1973).

The restoration of North American bison is considered as one of the first conservation success stories and is seen as a model of natural resource conservation (Ward 2000). To date there are approximately 500,000 bison in both private (raised as livestock) and conservation herds (Boyd 2003). Nearly all modern plains bison are descendants of the 76 to 84 bison that were used to establish the 5 private bison herds that aided in the recovery of American bison in the 1800s, along with the wild bison population in Yellowstone National Park (Garretson 1938; Meagher 1973; Coder 1975). Bison overcame multiple historic climatic periods with extreme temperature, moisture and ecological changes, imported parasitic, bacterial and viral diseases from Europe and Africa, and widespread habitat destruction and population fragmentation and still continue to thrive. Nevertheless, effective genetic management strategies are still

required to ensure their long-term conservation due to widespread genetic contamination from domestic cattle and the potential loss of genetic diversity caused by multiple population bottlenecks and extreme population fragmentation (Halbert 2003).

Current Studies for North American Bison

Microsatellite studies have found allele frequency differences between some herds of wood bison and plains bison, but all current wood bison populations have been shown to contain genetic material from plains bison (Cronin *et al*. 2013). Douglas *et al.* (2011) examined the complete mitochondrial DNA sequences of wood buffalo and plains bison, and found that the two wood bison haplotypes were not monophyletic; instead they were inter-mixed with the other 15 bison haplotypes. Therefore, current populations of *B. bison bison* and *B. bison athabascae* are not significantly different with respect to their mitochondrial genomes. Cronin *et al*. (2013) concluded that the subspecies ranking of plains and wood bison was not supported by phylogenetic distinction and could be considered a northwestern (geographic) subpopulation of North American bison, fueling the debate that wood buffalo and plains bison are not genetically valid subspecies (Douglas *et al.* 2011).

The bison population bottleneck of the late 1800s may have occurred over such a short time period that significant genetic erosion was prevented. In fact, the nuclear genetic variation reported from modern bison is generally much greater than that of other mammalian species that have gone through similar population bottlenecks (Freese *et al.* 2007). With little information known on the genetic variation of bison before the population bottleneck, there is no way of knowing how much genetic variation was

captured by the herds that were rescued in the late 1800s (Halbert 2003). Genetic diversity within and between historic populations of plains bison across pre-colonized North America cannot be determined due to in large part of individuals moving between founding herds to help repopulate bison herds, making it hard to reconstruct historic genetic patterns (Freese *et al*. 2007). Levels of genetic variation in 11 federal bison herds were examined and the majority of genetic contribution was found to be contained within only 4 federal populations, National Bison Range (Montana), Wind Cave National Park (South Dakota), Yellowstone National Park (Wyoming, Montana, Idaho) and Wichita Mountains National Wildlife Refuge (Oklahoma; Halbert 2003). The genetic variation was found to be unevenly distributed among these 11 federal bison herds, and management of these herds must be done with consideration in order to conserve the long-term integrity of the bison genome (Halbert and Derr 2008).

Approximately 30,000 of the 500,000 bison in North America are found in conservation herds (Halbert *et al.* 2005). Genetic differences have been reported among the conservation herds of plains bison in North America (Freese *et al*., 2007). These herds, as well as those found to have introgression with domestic cattle, have distinct genetic composition compared with other bison populations due to unique bison alleles and allelic distributions found only in certain herds (Halbert 2003). While the primary focus for conservation of bison should be on those herds that are determined to be free of domestic cattle introgression, however, bison herds with low levels of domestic cattle introgression can have value for a conservation herd due to their historical importance and unique genetic makeup (Freese *et al*. 2007).

Current technologies can test for domestic cattle genetics, in both the mitochondrial (Polziehn *et al.* 1995; Ward *et al.* 1999) and nuclear DNA (Ward 2000; Halbert 2003; Halbert *et al.* 2005) in modern bison samples. Halbert *et al.* 2005 used 100 microsatellites that represented regions on 29 of the 30 bison chromosomes, as well as the X chromosome, that were available from the domestic cattle genome map databases. From these 100 microsatellites, 14 were determined to be used as a diagnostic test for introgression in bison that cover 7 genomic regions with 1.2 to 7.4 megabases in the bison genome with confirmed cattle introgression (Halbert *et al.* 2005). While these technologies and studies have been useful for detecting introgression within herds (e.g., >100 bison), they do not provide the needed resolution to detect cattle introgression in individual bison at the genomic level.

Over the past decade, many new genetic tools have been developed and applied to bison for both population management (e.g., parentage testing) and conservation efforts (e.g., assessment of relationships among populations and detection of introgression; Ward *et al.* 1999; Schnabel *et al.* 2000; Halbert *et al.* 2007). These currently available genetic technologies for bison management lack the resolution and coverage of the bison genome that whole-genome sequencing offers. Whole-genome sequencing provides the next step in advancing bison management and conservation; however, a reference bison assembly is currently not available. Although the cattle genome sequence is available, using it as a guide to assemble a bison reference sequence would create domestic cattle reads in the bison sequence and lead to inconsistent alignments, misplaced reads while comparing sequences, and will not reflect all of the

novelty of the bison genome (Gnerre *et al.* 2009). Therefore, a *de novo* bison reference assembly will allow for an unbiased genomic sequence determination for the North American bison.

Whole Genome Sequencing as a Solution

Whole-genome sequencing offers new technologies that will advance bison management and conservation at a deeper genomic level than current technologies can offer. Combined next generation sequencing platforms of both long (454) and short (Illumina) DNA sequence reads provided a deeper and more complete *de novo* reference bison genome sequence that will not rely on the domestic cattle reference for assembly and annotation and be unbiased. If the domestic cattle sequence was used to complete the bison reference sequence it would limit accurate sequence comparison between bison and domestic cattle and may result in miscalled variants between the two genomes. The opportunities now available for utilizing genomic technology offers insight into the genetic history, taxonomy and inheritance of genetic traits in bison as never before possible.

De Novo *Bison Reference Sequence*

The bison population at Yellowstone National Park is one of the most thoroughly studied and most well-known of the public bison herds in North America. Ward *et al.* (1999) found no evidence of domestic cattle mitochondrial DNA and 2 distinct haplotypes at Yellowstone National Park, as well as no detection of nuclear introgression of domestic cattle (Ward *et al.* 2000; Halbert 2003). Therefore, our bison reference

animal is a well-documented male plains bison from Yellowstone National Park designated as “Templeton” after the late geneticist Dr. Joe W. Templeton.

The *de novo* bison reference sequence was assembled utilizing both mate-pair and pair-end read technologies to give approximately 75X coverage across the entire genome. The bison reference assembly provided a way to detect genetic variants in bison populations, including single nucleotide variants (SNVs) and insertion and deletions (INDELs) at the genomic level. A comparison of the bison and domestic cattle reference genomes allowed identification of interspecific genomic variations and their associated genes. Whole genome sequencing technologies provides a better tool to detect introgression of cattle genetics into the bison genome and more in-depth genomic analysis to be used for bison conservation management.

Additional/Whole Genome Re-sequencing Candidates

Samples for re-sequencing by paired-end technologies were chosen to represent wood buffalo, southern plains bison, Yellowstone National Park, and historical bison. A pure sub-population of wood buffalo (Banfield and Novakowski 1960) was believed to have been used to establish populations at Mackenzie Bison Sanctuary and Elk Island National Park in Canada (Geist 1991). Therefore, 4 male, wood buffalo samples were chosen to be sequenced to represent the wood buffalo population from Elk Island National Park. In addition, the last remaining 36 bison from Charles Goodnight’s herd were used to create the Texas State Bison Herd (TSBH) after being relocated to Caprock Canyons State Park in 1997 (Swepston 2001). Therefore, the only modern descendants of the southern plains bison are believed to be found in the Texas State Bison Herd and

four male bison from Caprock Canyons State Park were chosen to be candidates for whole genome re-sequencing (Halbert 2003). Four additional Yellowstone National Park female bison were chosen to be re-sequenced in order to develop a deeper understanding of the genomic diversity in this historically important bison population. These animals were documented from two separate collection times. Two samples were collected in the year 2000, while the other 2 were collected in the year 2009. In addition, in order to provide a historical context to this study, bison samples were collected and DNA was isolated from museum specimens housed at the Smithsonian Museum of Natural History (Smithsonian Institution) in Washington D.C. Selected historic bison samples were all well documented by museum records and collected from 1850 to the1880s. Two samples were chosen for whole genome sequencing; a female skull sample (015696; designated historical sample 6 (S6)) that was collected November 3, 1886 by Dr. William Hornaday, in Dawson County, Montana and bone samples associated with a bison skull (002007; designated historical sample 9 (S9)) that was collected in August 1856 by Dr. Ferdinand Vandeveer Hayden in the area that would become the Hayden Valley in Yellowstone National Park.

With the completion of the *de novo* plains bison reference genome sequence we compared the re-sequenced historic and modern bison sequences and identified genomic variants and their associated functional genes. This multi-way comparison of bison genome sequences was used to determine genomic differences between each population and the reference population, as well as comparing these differences across populations. These variants were used to evaluate evolutionary differences between modern and

historical bison, and to help determine taxonomic status of bison sub-species. In addition, by directly comparing whole genomes between domestic cattle and bison reference sequences, it is possible to define differences between these two closely related species and provide a foundation for developing an extremely robust test for introgression in bison.

Study Objectives

The objectives of this study were to provide approximately 75X coverage *de novo* bison reference sequence assembly of a North American plains bison for characterization of genetic variation. A comparative genomic analysis was performed among modern bison, historic bison samples, and domestic cattle and identified genomic variants that can detect introgression of the domestic cattle genome into modern bison. Identified genomic variants were then annotated to determine the effects on gene structure, phenotypic traits and fitness in bison.

With the completed *de novo* plains bison reference genome sequence assembly, comparisons of historic and modern bison sequences identified genomic variants and were compared across bison populations. Historic bison samples that predate introgression were sequenced and conserved genetics between historic and current bison were identified. Identified variants between modern and historic bison provided an outline of the genetic architecture of bison that existed before the population bottleneck and a more in-depth genomic analysis that will be used for bison conservation management of populations of bison.

CHAPTER II

DE NOVO ASSEMBLY AND ANNOTATION OF THE NORTH AMERICAN BISON (*BISON BISON*) REFERENCE GENOME

Introduction

American bison *(Bison bison*) are a well-known iconic species that symbolize the early colonization of western North America. Unfortunately, the American colonization of North America in the late 1800's resulted in the almost complete decimation of the American bison and subsequent population bottleneck (Coder 1975; Dary 1989). Bison were also faced with forced hybridization, or introgression of domestic cattle genetics *(Bos taurus),* through the failed experiments of some ranchers to produce a hardier beef animal for the great-plains (Coder 1975). The hybridization of domestic cattle into bison presents challenges in the management and conservation of the American bison today, primarily because it is difficult to differentiate between hybrid cattle-bison and purebred bison within a population (Douglas *et al.* 2011). With little information known on the genetic variation of bison before the population bottleneck, we cannot ensure the level of genetic variation captured by the herds rescued in the late 1800s (Halbert 2003).

Over the past decade, many new genetic tools have been developed and applied to bison for both population management (e.g., parentage testing, identification of loci for commercially important traits) and conservation efforts (e.g., assessment of relationships among populations, detection of introgression; Polziehn *et al.* 1996; Ward *et al.* 1999; Schnabel *et al.* 2000; Halbert *et al.* 2007). Recent studies demonstrated

many, but not all, bison herds have traces of cattle DNA as a result of this hybridization (Polziehn *et al.* 1995; Ward *et al.* 1999, 2000; Halbert *et al.* 2005). While these technologies are useful for detecting introgression within herds (e.g., >100 bison), they do not provide the needed resolution to detect cattle introgression in individual bison at the genomic level.

The bison population at Yellowstone National Park (YNP) is one of the most thoroughly studied and well-known public bison herd in North America (Halbert 2003). Ward *et al.* 1999 found no evidence of domestic cattle mitochondrial DNA and 2 mitochondrial DNA haplotypes at YNP. Twenty-eight YNP bison were analyzed for 21 microsatellites used for the detection of domestic cattle nuclear introgression in bison and no signs of introgression were detected (Ward *et al.* 2001). This was supported by Halbert 2003 who reported that YNP was found to have high levels of genetic variation and no detected domestic cattle introgression from 488 animals. Therefore, the animal chosen to provide the bison genome reference sequence was an adult male from YNP.

Whole-genome sequencing provides the next step in advancing bison management and conservation; however, a reference bison assembly was not previously available to utilize the genomic capabilities that whole-genome sequencing can offer researchers. Whole genome sequencing assemblies are usually comprised of contigs and scaffolds. Contigs are overlapping genome sequences where the orders of bases are known at a high confidence level, while scaffolds are longer and comprised of contigs and gaps. Most genome assemblies cannot be placed onto chromosomes due to insufficient mapping information and only reach the scaffold level. The scaffold N50 is

a good estimate of accurate assembly, with a longer scaffold N50 representing a better assembly. A 2.82-Giga-base *de novo* reference assembly of the American bison genome was produced using approximately 75X coverage, utilizing both mate pair and paired-end sequencing, comprising of 128,431 scaffolds and 470,415 contigs. The N50 scaffold for this assembly is 7,192,658 base pairs.

Raw (unassembled) sequence reads from Illumina (short reads) and 454 (long reads) were mapped to both nuclear and mitochondrial sequences of the domestic cattle reference UMD3.1 (Ensembl GCA_000003055.3), in order to detect genetic variants, including single nucleotide variants (SNVs) and insertion and deletions (INDELs). These variants were quantified and annotated in order to examine their effect on gene structure and function in bison. Biological processes enriched for these variants were analyzed and compared between bison and domestic cattle.

Comparison of the annotated reference bison genome and domestic cattle provides a resource that allows for the identification of genomic variations and their associated genes. This information, in turn, can be used to provide informed management and breeding of modern bison today. This reference bison assembly allows insight into the genetic history, taxonomy, and inheritance of important genetic traits in bison that have allowed them to thrive over the years.

Materials and Methods

Collection of DNA Samples/Isolation of DNA

Templeton (Figure 1) is a well-documented bison from YNP. He was a member of a brucellosis free herd that existed on the Green Ranch in Montana. Yellowstone

National Park bison are considered free from domestic cattle introgression. In March of 2011, blood, hair, and tissue samples were collected from this North American plains bison. DNA was isolated from 15 mL of blood by using a standard phenol-chloroform-isoamyl alcohol (PCI) extraction protocol (Sambrook *et al.* 1989).

Figure 1. The bison reference genome animal Templeton.

Microsatellite and Mitochondrial Genotype Analysis

All of the current technologies available in our lab for domestic cattle introgression (14 nuclear and mitochondrial) and an additional twenty-six polymorphic markers were genotyped from the reference animal prior to the sequencing of its genome to ensure that the selected sample did not have detectable domestic cattle introgression (Ward *et al.* 1999; Schnabel *et al.* 2000; Halbert *et al.* 2003). PCR reactions consisted of 5 µL total volume with: 1 µL of DNA (extracted from hair follicles described by KAPA

Express Extraction Kits, KapaBiosystems); 0.05 to 0.4 µM each primer; 1x MasterAmp PCR Enhancer (Epicentre, Madison, Wisconsin); 500 µM deoxynucleotide triphosphates 3.0 mM $MgCl_2$, 1x reaction buffer; 0.5 units *Taq* DNA polymerase (Promega, Madison, Wisconsin). PCR products were separated on an ABI 3130 Genetic Analyzer (Applied Biosystems, Foster City, California) using an internal size standard (Mapmarker 400, Bioventures, Inc., Murfreesboro, Tennessee). GeneMapper 3.7 software (Applied Biosystems) was used for allele identification and comparison. Excel Microsatellite Toolkit (Park 2001) was used to obtain values for heterozygosity and average number of alleles per locus for 26 of the polymorphic markers for Templeton. The relationship of Templeton to the 8 core U.S. federal bison herds was assessed using the multilocus Bayesian clustering method across 10 iterations, with *K* (number of known populations) set equal to 8 in the program Structure 2.1 (Pritchard *et. al* 2000; as described by Halbert and Derr 2008). List of 40 loci used and the genotypes can be found in Table 2.

Karyotyping

Karyotyping was performed to ensure normal chromosomes were obtained from Templeton. Sodium heparin-stabilized peripheral blood was used for Pokeweed-stimulated short-term lymphocyte cultures, followed by metaphase chromosome preparations that were done according to standard cytogenetic methods (Raudsepp and Chowdhary 2008). Standard Giemsa staining with 5% Geimsa stain and G-banding cells of chromosome spreads were performed (Seabright 1972). In total, twenty cells were captured and analyzed and 4 Giemsa stained and 4 G-banded cells karyotyped. Slides

were examined and images were captured using a Zeiss Axioplan2 fluorescent microscope, and metaphases were analyzed with Ikaros (MetaSystems GmbH) software.

Additionally, fluorescence *in-situ* hybridization (FISH) was done using domestic cattle bacterial artificial chromosome (BAC) clones from the bovine TAMBT BAC library (Cai *et al.* 1995) containing two pseudoautosomal genes *GYG2* (clone 235H1) and *CRLF2* (clone 7138-24C10) used to confirm the presence of the X and Y chromosomes. BAC DNA was isolated using Plasmid Midi Kit (Qiagen) according to the manufacturer's instructions and labeled by nick translation with digoxigenin-11-dUTP following standard protocols (Raudsepp and Chowdhary 2008). Probes were hybridized to bison chromosomes, and the signals were detected with anti-dig-Rhodamine (red). Altogether, 10 metaphase cells were analyzed and images captured with a Zeiss Axioplan2 fluorescent microscope and Isis V5.2 (MetaSystems GmbH) software.

Whole-genome Sequencing, Assembly and Annotation

The American Bison genome was sequenced using a *de novo* assembly method that utilizes hybrid Illumina and 454 sequencing data. Using 30 micrograms of genomic DNA (from the above extracted DNA), sequencing libraries were generated for 4 20kb

paired-end single stranded libraries for sequencing on a 454 GS-FLX Titanium™ sequencer following manufacturers protocol (GS FLX Titanium Series; Roche Applied Sciences) and were circularized using 'titanium' 42Bp linker at United States Department of Agriculture Meat and Animal Research Center by Dr. Tim Smith and Rene Godtel. 454 mate pair libraries were also constructed and ran on the 454 at Iowa State University by Dr. David Alt.

Libraries were also prepared following manufacturers protocols for shotgun sequencing of 10 paired-end libraries with approximate 390 base pair insert size and a 5-kb Nextera jump mate-pair library for sequencing on the Illumina HiSeq 2000™ Next-Gen using the 100 cycle paired-end normal mode from the above extracted DNA (Illumina, San Diego, CA). DNA sequences were generated by the Illumina sequencing machine at Iowa State University. Statistics for the combined Illumina paired-end files and mate pair, as well as combined 454 paired-end files can be found in Table 1. The library mean size is the mean average in base pairs for each library that was sequenced and that libraries standard deviation for all of the libraries produced for each sequencing technology (Table 1).

Table 1. Statistics for Illumina (paired-end and mate pair) and 454 paired-end libraries used for *de novo* bison reference sequence.

Library	Average read length	Number of reads (Millions)	Library mean size (base pairs)	Library standard deviation (base pairs)
Illumina:				
Paired-end	101	1115	300	40
Mate pair	101	85	4000	800
	101	239	4500	900
	101	531	6000	1000
454:				
Paired-end	398	25.6	15000	3500

DNA sequence files were used to produce approximately a 75X coverage of a *de novo* reference assembly. The reference assembly was performed in collaboration with the University of Maryland and Dr. Aleksey Zimin using the MaSuRCA assembler version 2.0.4 (Zimin *et al.* 2013). The MaSuRCA assembler is based on the idea of using a combination of the de Bruijn graph and the Overlap-Layout-Consensus (OLC; Celera Assembler version 6.1) methods. This is achieved by reducing the most numerous and high coverage Illumina paired-end reads to a much smaller set of long consensus super-reads. The super-reads are then assembled using the OLC method along with the error corrected and filtered Illumina linking mate pair reads and the 454 paired-end reads. For

the American bison genome the high-coverage paired-end Illumina reads data were reduced to 7.2 billion bases spread across 26.7 million super-reads with an average length of 269 bases. Utilization of the super-reads reduces the problem of assembling millions of short reads down to a 100 times smaller. The MaSuRCA assembly contained 2.76 billion bases of sequence in scaffolds with an N50 contig size of 18,824 bases and N50 scaffold size of 6.8 million bases.

This whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession number JPYT00000000. The version described in this paper is version JPYT01000000.

Annotation of the *de novo* bison reference genome sequence was completed using the assembled bison reference sequence and RNA sequences by the National Center for Biotechnology Information (NCBI; Thibaud-Nissen *et al.* 2013). The GenBank assembly accession is GCA_000754665.1 and RefSeq assembly accession is GCF_000754665.1 for UMD1.0.

RNA Sample Collection

Select tissues were obtained from Bison #423 (organism = *Bison bison*), a 3-year old healthy cow (USDA National Animal Disease Center, Ames, IA). Upon euthanasia (Fatal-Plus, Vortech Pharmaceuticals, LTD., right jugular vein), 100mg tissue samples were immediately harvested, flash-frozen in liquid nitrogen, and stored at -80^{o}C until RNA purification was performed.

RNA Purification

Purification of total RNA from frozen tissues was performed in an identical fashion using TRIzol© and the PureLink RNA Mini Kit (Ambion Life Technologies, Cat. No. 15596-018 and 12183025, respectively). 100mg frozen tissue samples were homogenized in (l)N_2 using pre-chilled mortars and pestles. Pulverized tissues were immediately suspended in 1mL room-temperature TRIzol © and transferred to 1.5mL RNase-free Eppendorf Microcentrifuge tubes on ice. Samples were centrifuged at $\geq$ 12,000 x *g* for 10 minutes at 4ºC and cleared supernatants recovered. Phase separation was performed by the addition of 0.2mL chloroform per 1mL TRIzol © Reagent used for homogenization followed by vigorous hand mixing and centrifugation at $\geq$ 12,000 x *g* for 15 minutes at 4ºC. An equal volume 100% ethanol was added to the retained supernatant and vortexed well. Samples were transferred to PureLink spin cartridges with collection tubes and centrifuged at $\geq$ 12,000 x *g* for 1 minute at room temperature. Flow-through was discarded and 500µL Wash Buffer II was added to the spin cartridge and centrifuged at $\geq$ 12,000 x *g* for 15 seconds at room temperature and repeated a second time followed by a final centrifugation at $\geq$ 12,000 x *g* for 1 minute at room temperature to dry the membrane with bound RNA. Bound RNA was eluted by the addition of 100µL RNase-free H_2O to the center of the spin cartridge and recovery tube and centrifugation for 2 minutes at $\geq$ 12,000 x *g* at room temperature. Eluted samples were stored at -80ºC. Total RNA purification was verified using NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific) and 2100 BioAnalyzer and RNA 6000 Nano Chip (Agilent Technologies). Total RNA concentrations and RNA integrity

numbers (RIN) ranged from 36-166ng/µL and 7.1-7.9, respectively. 2µg total RNA (in 10µL DEPC-treated H_2O) from liver, spleen, lung, skeletal muscle, kidney and supramammary lymph node tissues were provided to Iowa State University (Ames, IA) for downstream rRNA reduction, library construction and sequencing. Directional libraries were prepared according to manufacturer's directions using the TruSeq Stranded Total RNA sample prep kit with human/mouse/rat RiboZero (Illumina, Inc.) treatment for removal of rRNA prior to library preparation.

Sequence Alignment

Both the paired-end and mate-pair sequences of the bison reference raw reads were trimmed using FASTQ-MCF filtering out bases with a quality score less than 20 from each individual read and reads with a remaining sequence length of less than 70 bases (Aronesty 2011). Whole Systems Genomics Initiative (WSGI) provided the computational resources and systems administration support for the WSGI HPC Cluster used for this analysis. The filtered reads were then aligned to the domestic cattle UMD3.1 reference sequence using Burrows-Wheeler Alignment version 0.6.2 (BWA-MEM; Li 2013) using the default settings. The resulting BAM (binary short DNA sequence read alignment; Li *et al.* 2009) files were combined using the merge option of the Sequence Alignment/Map (SAM)tools 0.1.18 software package (Li *et al.* 2009). Read group information was added using the AddOrRelpaceReadGroups option of PicardTools 1.7.1 (https://github.com/broadinstitute/picard/releases/tag/1.128). Then Genome Analysis Toolkit 3.1.1 (GATK; McKenna *et al.* 2010) option RealignerTargetCreator was used to realign and account for INDEL shifted coordinates

to create a realigned and sorted BAM file of read alignments to UMD3.1 reference. Finally the SAMtools view and flagstat options (Li *et al.* 2009) were used to obtain statistics of the alignment of the bison reference genome to the domestic cattle reference genome.

Identification of Genetic Variants and Analysis

Genetic variants, SNVs and INDELs, were called against both the bison and the cattle references and were filtered according to the GATK Best Practices recommendations (DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). The resultant variants were placed into variant call formatted (VCF) files. VCFtools 0.1.11 vcf-stats (Danecek *et al.* 2011) option was used to determine basic statistics and counts of the SNVs and INDELs.

These identified variants were then annotated using the SnpEff 4.1 software (Cingolani *et al.* 2012) against the UMD3.1.76 reference from Ensembl. The annotated variants were then analyzed using the Database for Annotation, Visualization and Integrated Discovery (DAVID; Huang *et. al* 2009) version 6.7 Functional Annotation Tool (FAT; http://david.abcc.ncifcrf.gov/home.jsp) in order to identify enriched biological pathways.

Pseudo-Chromosome Mapping

Pseudo-chromosomes were produced using the UMD3.1.76 gff (http://useast.ensembl.org/Bos_taurus/Info/Annotation) chromosome file from Ensembl (Flicek *et al.* 2014) and scaffolds of bison reference sequence to create synteny blocks using the software Symap 4.2 (Soderlund *et al.* 2006).

Results

Microsatellite and Mitochondrial Genotype Analysis

Templeton was found to have bison mitochondrial DNA genotype and no domestic cattle introgression alleles were detected and alleles for microsatellites can be found in Table 2. Templeton's main genetic contribution when compared to the 8 core U.S. federal bison herds, was as expected, with 91.0% of his genome coming from Yellowstone Nation Park (Figure 2).

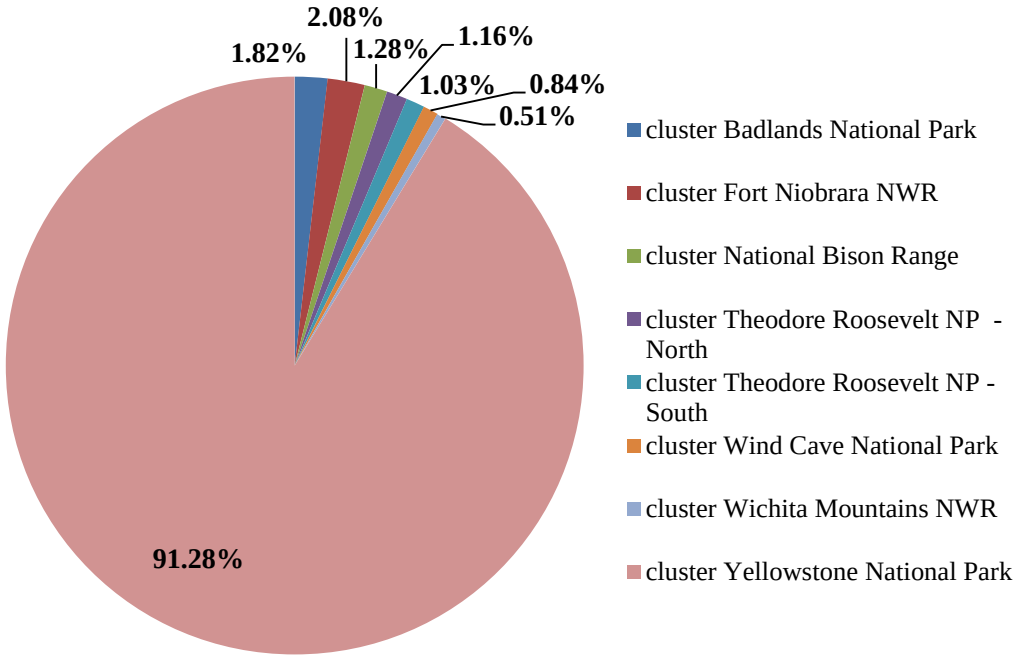


Figure 2. Genomic contribution (percentage) of 8 core U.S. federal bison herds to Templeton.

Table 2. List of loci used and genotypes for Templeton.

Locus	Allele 1	Allele 2
AGLA17	215	215
AGLA293	218	218
BL1036	191	191
BM1225	253	271
BM1314	137	137
BM1706	238	252
BM17132	85	85
BM1862	205	207
BM1905	176	176
BM2113	143	143
BM4107	165	183
BM4307	185	185
BM4311	92	98
BM4440	125	127
BM4513	132	132
BM47	103	103
BM6017	118	118
BM711	161	167
BM7145	108	108
BM720	213	235
BMS1001	115	115
BMS1074	158	160
BMS1315	135	135
BMS1675	87	87
BMS1716	189	191
BMS1857	150	158
BMS2270	68	68
BMS4040	75	75
BMS410	83	97
BMS510	91	94
BMS527	167	175
CSSM36	158	158
CSSM42	167	171
HUJ246	262	262
ILSTS102	147	147
INRA189	96	96
RM185	92	92
RM372	132	134
RM500	123	123
SPS113	130	132
TGLA122	140	148
TGLA227	73	73

Karyotyping

Templeton was found to have normal chromosomes and a diploid number of 2n=60 (Figure 3) and normal X and Y chromosomes (Figure 4). Cattle PAR was mapped to the short arm of the bison Y chromosome, metacentric, showing that the y chromosome is structurally different than the *Bos taurus* Y chromosome which is sub-metacentric (Di Meo et al. 2005).

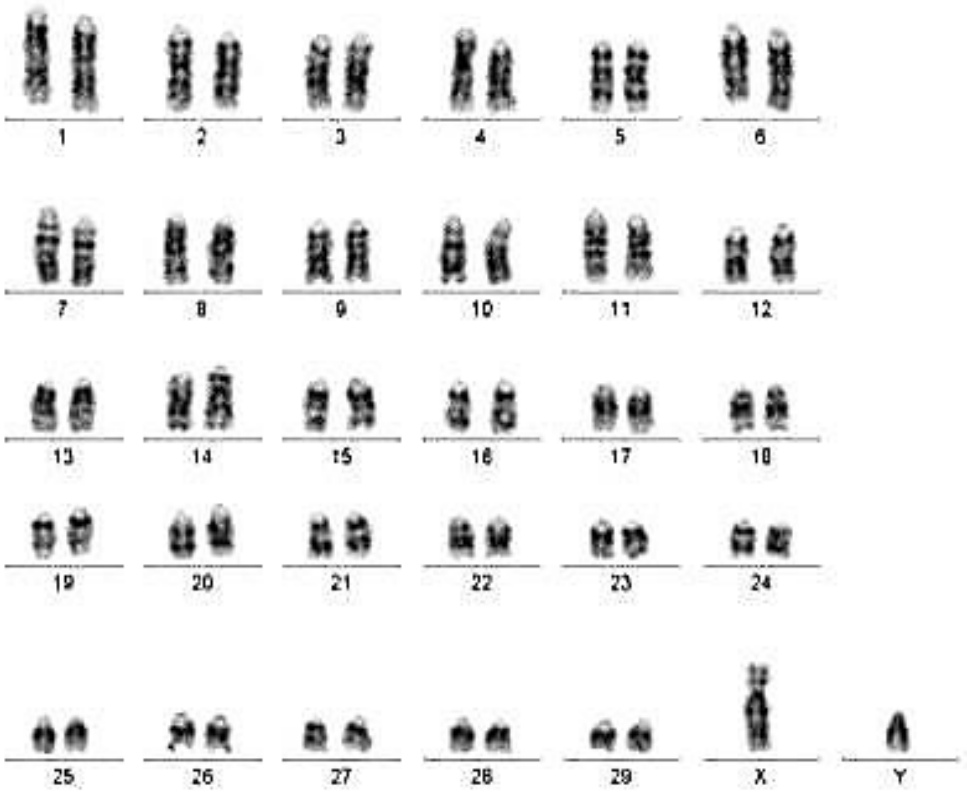


Figure 3. Templeton's G-banded karyotypes showing normal diploid chromosome number 2n=60.

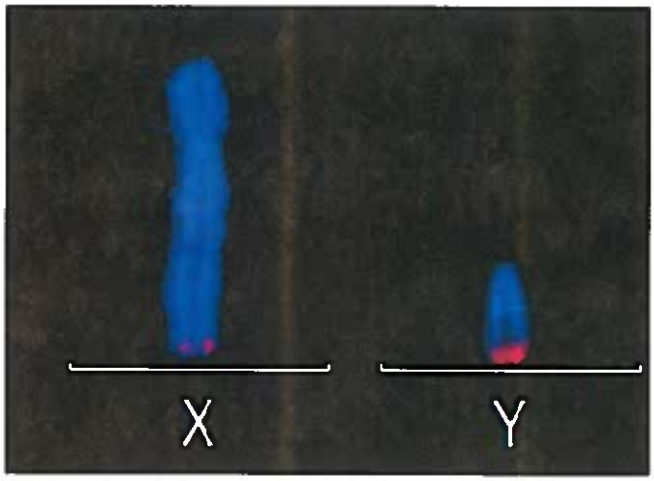


Figure 4. Identification to Templeton's X and Y chromosomes by FISH as marked by red fluorescence.

Annotation

Annotation of the bison *de novo* reference genome was done by NCBI, using the MaSuRCA generated files for contigs and scaffolds and the RNA sequence data. The RNA sequences generated were deposited into the NCBI Sequence Read Archive http://www.ncbi.nlm.nih.gov/sra under accession number SRX765353. This whole genome shotgun sequencing project has been deposited at DDBJ/EMBL/GenBank under the accession JPYT00000000 version JPYT00000000.1 GI:684690141. The annotation release, statistics and reports can be found at http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bison_bison_bison/100/%3Chttp://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bison_bison_bison/100/.

The bison genome reference assembly can be found with the assembly accession number GCF_000754665.1 and assembly name Bison_UMD1.0 at http://www.ncbi.nlm.nih.gov/assembly/GCF_000754665.1/. The database link can be found using the BioProject ID: PRJNA257088 and the BioSample ID: SAMN02947321 (NCBI).

The NCBI annotated bison reference genome, Bison_UMD1.0, contains approximately 2.82 Gigabases of total sequence length, and is composed of 128,431 scaffolds and 470,415 contigs. The scaffold N50 for Bison_UMD1.0 is 7,192,658 base pairs (Table 3), validating that our assembly is of good quality. Global statistics for the bison annotation can be found in Table 3. The annotation done by NCBI reported a count of 26,001 genes and pseudogenes (Table 4). 20,782 of the 26,001 (79.9%) genes were found to be protein-coding, and 6,154 were genes with variants, those genes that are represented by multiple, alternatively spliced transcript variants.

Table 3. Global statistics (in base pairs) for Bison_UMD1.0 (NCBI).

	Bison_UMD1.0
Total sequence length	2,828,031,685
Total assembly gap length	195,767,988
Gaps between scaffolds	0
Number of scaffolds	128,431
Scaffold N50	7,192,658
Number of contigs	470,415
Contig N50	19,971

Table 4. Gene and feature statistics (NCBI).

Feature	Bison_UMD1.0
Genes and pseudogenes	26,001
protein-coding	20,782
non-coding	1,677
pseudogenes	3,542
genes with variants	6,158

When compared with the domestic cattle (UMD3.1) and human reference genome annotations (both HuRef_1 and HuRef2 (GRCh38)) the bison reference total sequence length was slightly larger than the cattle annotation and smaller than the 2 human reference genome annotations (Table 5). The bison genome does have less genes and pseudogenes when compared to the other 3 annotations, but was found to have more protein coding genes than the others (Table 5).

Table 5. Bison (UMD1.0) reference genome annotation comparison to domestic cattle (UMD3.1) and human (HuRef_1 and HuRef2 (GRCh38)) reference genome annotations.

Feature	Bison_UMD1.0	Cattle_UMD3.1	HuRef_1	HuRef_2 (GRCh38)
Total sequence length (base pairs)	2,828,031,685	2,670,422,299	2,844,000,504	3,209,286,105
Total number of chromosomes and organelles	31	31	24	25
Genes and pseudogenes	26,001	26,740	39,480	41,722
protein-coding	20,782	19,994	19,691	20,246
non-coding	1,677	3,825	8,555	9,153
pseudogenes	3,542	797	11,234	12,323
genes with variants	6,158	2,581	9,563	14,632
mtDNA size	16,319	16,338		16,569

In total there were 15,397 genes and pseudogenes found to be in common with the bison, cow, and human genome annotations when comparing the Gene Symbols and descriptions from gene reports. A total of 5,325 genes and pseudogenes were found to only be in the bison and domestic cattle annotation based on gene symbol and description and not the human annotation gene list, and only 227 genes and pseudogenes were found to be in common with the bison and human annotations only. Lastly, 5,053 genes and pseudogenes were found to only be in the bison annotation when compared to the domestic cattle and human annotation gene lists based on both Gene Symbol and description. Of these 5,053 genes the main gene description was an endogenous retrovirus group K member 9 Pol protein-like gene, with 17 different genes having this description. Most of these genes found only in bison were pseudogenes or had similar function in humans and cattle. Future analysis will be needed to determine what these genes and pseudogenes functions are found to be in bison.

Sequence Alignment

The mem alignment option of BWA was used to align raw bison DNA sequence paired-end and mate pair reads totaling 1,008,038,624 reads, created by Illumina sequencing technology, to the UMD3.1 domestic cattle reference. The SAMtools options, view and flagstat, were used to obtain statistics of the bison Illumina paired-end reads mapped to the domestic cattle reference sequence. A total of 993,981,233 of the 1,008,038,624 (98.6%) bison reads were mapped to domestic cattle, with 944,493,355 (93.70%) reads properly mapped (Table 6). Unmapped reads are those that did not have a mate mapped and were excluded from alignment.

Table 6. Samtools flagstat statistics of the bison reference sequence mapped to UMD3.1 domestic cattle reference sequence.

Statistic	Reads (base pairs)
in total (QC-passed reads + QC-failed reads)	1,008,038,624
duplicates	0
mapped	993,981,233 (98.61%)
paired in sequencing	1,008,038,624
read1	503,120,843
read2	504,917,781
properly paired	944,493,355 (93.70%)
with itself and mate mapped	991,194,251
singletons	2,786,982 (0.28%)
with mate mapped to a different chr	43,008,301
with mate mapped to a different chr (mapQ>=5)	18,766,379

Identification of Genetic Variants and Analysis

SNVs and INDELs were called and identified separately using GATK against the bison and domestic cattle reference genomes. Samtools VCF-stats and SnpEff were used to determine basic statistics and counts of SNVs and INDELS for the bison variants detected. A total of 28,443,364 SNVs were discovered between Templeton and the domestic cattle reference, with 22,073,944 (approximately 77.6%) SNVs being homozygous for the variant allele (no SNV was from the reference; Table 7). Only 6,329,185 (approximately 22.3%) reference alleles of the 28,443,364 SNVs that were detected occurred when the variant was heterozygous for the reference (cattle) and the bison variant allele. There are some positions in the bison genome that are going to contain the same genomic sequences since they derived from a common ancestor 0.5-2 million years ago in Eurasia (McDonald 1981). There were 40,235 multi-allelic VCF entries, which means that Templeton was heterozygous at that position, but for 2

different variant alleles, not a reference allele. Overall there was one SNV detected every 93 bases and 32,086,858 genome region and coding effects found caused by the SNVs discovered. The most common variant between Templeton and domestic cattle was G>A substitution at 4,953,362 SNVs found, with the least common substitution found was A>T with only 969,768 detected (Table 8).

There were 2,627,645 INDELs discovered against both bison and domestic cattle, with 1,233,140 (46.9%) insertions and 1,394,505 (53.1%) deletions. All INDELs were annotated and 29,940 were classified as multi-allelic VCF entries (Table 7). There were 2,976,475 effects detected by SnpEff from these INDELs, with a variant rate of 1 variant every 1,012 bases. Chromosomal variant counts for both SNVs and INDELs for bison onto domestic cattle can be found in Figure 5, with chromosome 1 having the most detected variants. Figure 6 shows the count of variants with corresponding quality scores of the SNVs and INDELs annotated with SnpEff after filtering. This helps to verify that low quality variants were not included in the downstream analysis and were properly removed.

Table 7. Summary statistics for SNVs and INDELs found in Templeton compared to domestic cattle.

	SNVs	INDELs
Warnings	2,898,711	271,063
Errors	173,652	6,352
Number of lines (input file)	28,443,364	2,598,155
Number of variants (before filter)	28,483,599	2,627,645
Homozygous for variant allele	22,073,944	2,208,623
Heterozygous (one Reference one variant)	6,329,185	360,038
Reference Alleles	6,329,185	360,038
Number of multi-allelic VCF entries	40,235	29,494
Number of effects	32,086,858	2,976,475
Genome total length	2,670,424,944	2,670,423,585
Genome effective length	2,660,909,050	2,660,907,691
Variant rate	1 variant every 93 bases	1 variant every 1,012 bases

Table 8. Base changes (SNVs) between Templeton and Domestic cattle.

	A	C	G	T
A	0	1,256,272	4,738,442	969,768
C	1,219,311	0	1,117,609	4,929,151
G	4,953,362	1,120,805	0	1,222,449
T	974,191	4,720,601	1,261,638	0

Summary of within genome and region consequences for SNVs and INDELs with SnpEff provides the number of effects by type and region within the genome (Tables 9, 10 and 11), with 64.14% of the SNV effects being found in the intergenic region of domestic cattle, with the next highest being 25.99% as an intron variant. Of the 32,086,858 genomic effects identified 31,809,534 (99.14%) were found to be modifiers, and only 3,512 (0.011%) were found to have a high genomic impact. The majority (approximately 60.93%) of the genomic effects were found to be in the silent functional class, but approximately 38.51% were found to be missense effects. These genomic effects can be further examined in future research to fully understand what these effects can be controlling within the bison genome.

There were 34,751,094 transitions (Ts) and 15,806,449 transversions (Tv) detected. For whole genome studies a Ts/Tv ration is expected to be between 2-2.1, suggesting that few false positives generated by random sequencing errors were within the sequence; our Ts/Tv ratio of 2.1985 for SNVs confirms that we were able to properly detect variants with few false positives between bison and domestic cattle (Li 2011).

These identified SNVs and INDELs were annotated against the UMD3.1.76 reference from Ensembl to give Gene IDs that can be used for biological pathway analysis. The gene lists for SNVs and INDELs that were annotated were combined and 24,551 genes were identified by SnpEff. There were 8 biological types found to be associated with the genes annotated from the SNVs in bison. The main biological type for the annotated genes found for SNVs and INDELs was protein coding with 19,960 of

the 24,551 genes and 19,920 of the genes, respectively (Table 12). 54 of the gene IDs annotated from the SNVs were not in the gene list from the annotated INDELs.

DAVID was chosen to do a Gene Ontology (GO) analysis and to determine enriched biological pathways due to the accessibility of software on-line and comprehensive gene lists. The set of gene IDs that contained 24,551 annotated genes from SnpEff was found to have a match of 18,992 DAVID IDs in the *Bos taurus* database. Choosing only the gene ontology for biological pathway FAT option and using the Functional Annotation Chart, 48 enriched gene ontology categories for biological pathways were produced with a False Discovery Rate (FDR) P-value ≤ 0.05 comprising of 7,332 Ensembl Genes (Appendix A). These 48 GO terms were mainly associated with regulatory functions in domestic cattle, which can be used to examine the impact these genes have on regulation in bison.

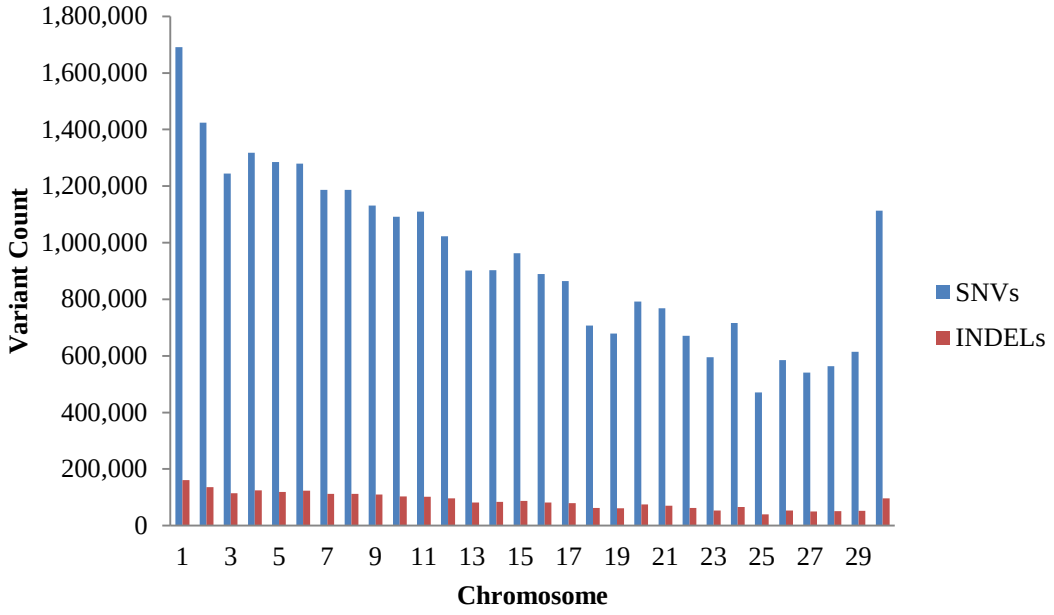


Figure 5. Variant (SNVs and INDELs) counts found for each chromosome from Templeton aligned to domestic cattle.

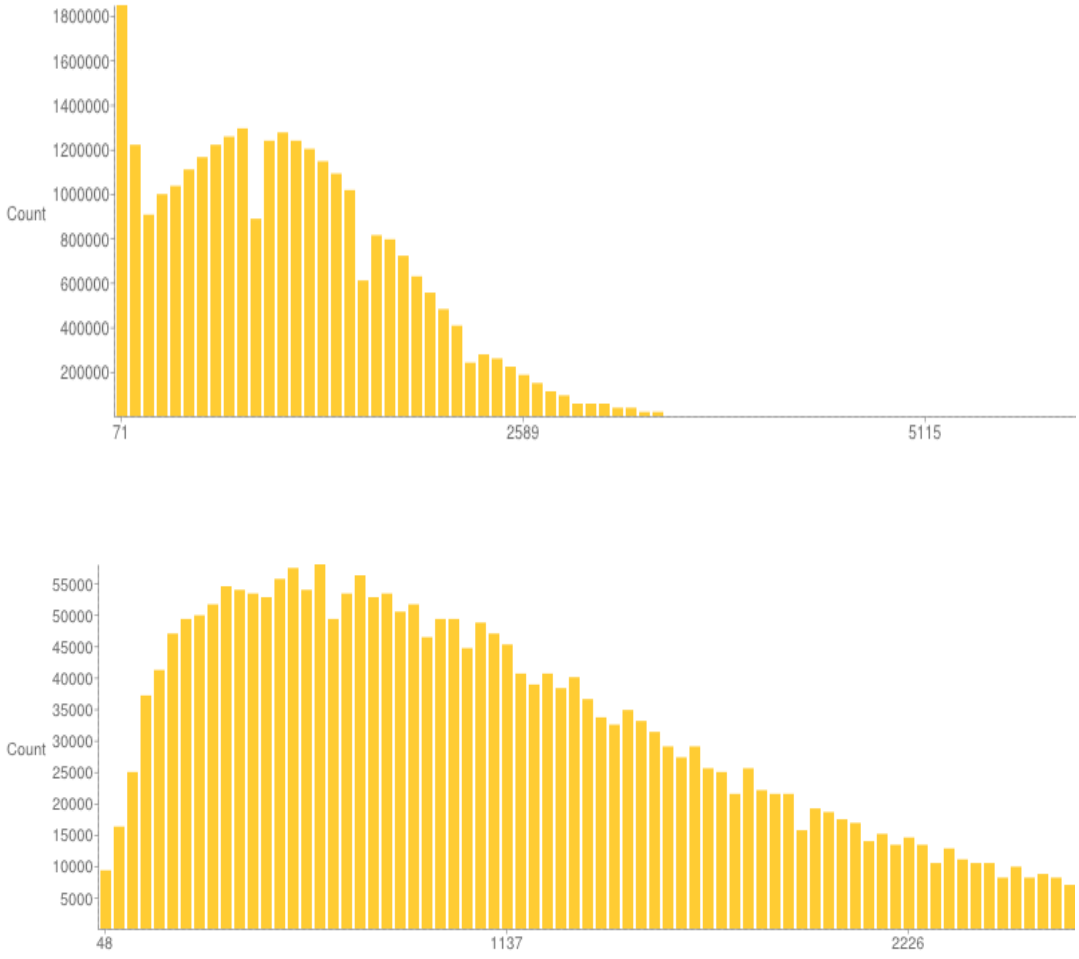


Figure 6. Variant counts with corresponding quality scores of SNVs (top) and INDELs (bottom) to evaluate the quality of variants annotated in Templeton.

Table 9. Number of consequences (effect type) in genome of bison found after SNVs annotated between bison and domestic cattle.

Effect Type	Count	Percent
None	173,652	0.54%
3_prime_UTR_variant	66,181	0.21%
5_prime_UTR_premature_start_codon_gain_variant	1,615	0.01%
5_prime_UTR_variant	12,031	0.04%
downstream_gene_variant	13,28,228	4.14%
initiator_codon_variant	6	0.00%
initiator_codon_variant+non_canonical_start_codon	2	0.00%
intergenic_region	20,581,948	64.14%
intron_variant	8,339,401	25.99%
missense_variant	95,501	0.30%
missense_variant+splice_region_variant	2,241	0.01%
missense_variant+splice_region_variant+splice_region_variant	4	0.00%
non_coding_exon_variant	12,901	0.04%
splice_acceptor_variant+intron_variant	524	0.00%
splice_acceptor_variant+splice_donor_variant+intron_variant	69	0.00%
splice_acceptor_variant+splice_region_variant+intron_variant	42	0.00%
splice_donor_variant+intron_variant	1,223	0.00%
splice_donor_variant+splice_region_variant+intron_variant	47	0.00%
splice_region_variant	574	0.00%
splice_region_variant+intron_variant	18,706	0.06%
splice_region_variant+non_coding_exon_variant	222	0.00%
splice_region_variant+splice_region_variant+intron_variant	25	0.00%
splice_region_variant+splice_region_variant+synonymous_variant	1	0.00%
splice_region_variant+stop_retained_variant	23	0.00%
splice_region_variant+synonymous_variant	3,153	0.01%
start_lost	74	0.00%
stop_gained	1,370	0.00%
stop_gained+splice_region_variant	63	0.00%
stop_lost	50	0.00%
stop_lost+splice_region_variant	50	0.00%
stop_retained_variant	60	0.00%
synonymous_variant	151,679	0.47%
upstream_gene_variant	1,295,192	4.04%

Table 10. Number of consequences (effect type) in genome of bison found after INDELs annotated between bison and domestic cattle.

Effect Type	Count	Percent
3_prime_UTR_variant	7,594	0.26%
5_prime_UTR_variant	697	0.02%
disruptive_inframe_deletion	279	0.01%
disruptive_inframe_deletion+splice_region_variant	8	0%
disruptive_inframe_insertion	189	0.01%
disruptive_inframe_insertion+splice_region_variant	6	0%
downstream_gene_variant	133,884	4.50%
frameshift_variant	1,382	0.05%
frameshift_variant+splice_acceptor_variant+splice_region_variant+intron_variant	11	0%
frameshift_variant+splice_acceptor_variant+splice_region_variant+splice_region_variant+intron_variant	13	0%
frameshift_variant+splice_donor_variant+splice_region_variant+intron_variant	8	0%
frameshift_variant+splice_donor_variant+splice_region_variant+splice_region_variant+intron_variant	10	0%
frameshift_variant+splice_region_variant	261	0.01%
frameshift_variant+start_lost	10	0%
frameshift_variant+stop_gained	20	0.00%
inframe_deletion	152	0.01%
inframe_deletion+splice_region_variant	5	0%
inframe_insertion	167	0.01%
intergenic_region	1,882,377	63.24%
intron_variant	817,671	27.47%
non_coding_exon_variant	677	0.02%
none	6,352	0.21%
splice_acceptor_variant+intron_variant	53	0.00%
splice_acceptor_variant+splice_donor_variant+intron_variant	151	0.01%
splice_acceptor_variant+splice_region_variant+intron_variant	50	0.00%
splice_donor_variant+intron_variant	44	0.00%
splice_donor_variant+splice_region_variant+intron_variant	44	0.00%
splice_region_variant	59	0.00%
splice_region_variant+intron_variant	1,939	0.07%
splice_region_variant+non_coding_exon_variant	78	0.00%
transcript	20	0.00%
upstream_gene_variant	122,217	4.11%

Table 11. Genomic regions associated with annotated SNVs and INDELs found in bison.

	SNVs		INDELs	
Region Type	Count	Percent	Count	Percent
DOWNSTREAM	1,328,228	4.14%	133,884	4.50%
EXON	263,919	0.82%	3,213	0.11%
INTERGENIC	20,581,948	64.14%	1,882,377	63.24%
INTRON	83,39,401	25.99%	817,671	27.47%
NONE	173,652	0.54%	6,375	0.21%
SPLICE_SITE_ACCEPTOR	635	0.00%	266	0.01%
SPLICE_SITE_DONOR	1,270	0.00%	96	0.00%
SPLICE_SITE_REGION	22,704	0.07%	2,083	0.07%
TRANSCRIPT	82	0.00%	2	0%
UPSTREAM	1,295,192	4.04%	122,217	4.11%
UTR_3_PRIME	66,181	0.21%	7,594	0.26%
UTR_5_PRIME	13,646	0.04%	697	0.02%

Table 12. Biological functions of genes associated with annotated SNVs and INDELs in bison.

Biological Type	SNVs	INDELs
Protein coding	19,960	19,920
Ribosomal RNA	401	399
Miscellaneous RNA	175	175
Small nucleolar RNA	846	845
Pseudogene	626	621
Processed pseudogene	171	169
Micro RNA	1,152	1,150
Small nuclear RNA	1,220	1,218
Total	24,551	24,497

Pseudo-Chromosome Mapping

Since no previous chromosome map is available for bison we do not have placements of genes on chromosomes. Since bison and domestic cattle shared a common ancestor and have the same number of chromosomes, we used the domestic cattle reference to generate pseudo-chromosomes to provide gene placements on chromosomes. Symap 4.2 (Soderlund *et al.* 2006) was used to produce a synteny alignment between Templeton's scaffolds and chromosomes from the UMD3.1.76 domestic cattle reference. Symap was able to create 447 synteny anchors and mapped a total of 414 scaffolds to the 29 autosomes and the X chromosome of domestic cattle. Appendix B offers Templeton's scaffolds sorted by chromosome placements, synteny block assigned, scaffold start and end position, and domestic cattle start and end position. Synteny blocks (in black) anchored to domestic cattle (in grey) for all chromosomes can be found in Figure 7. Chromosome 1 was found to have the most scaffolds mapped to it with 30 synteny blocks anchored and can be viewed in Figure 8, while chromosome 26 was found to have the least amount (6) scaffolds placed on it (Table 13). In total, Templeton's scaffolds covered approximately 2,283,389,917 (85.5%) Gigabases of the 2,670,424,944 Gigabases UMD3.1.76 cattle reference. Even though these are different species they do contain similar chromosomal arrangements and gene placement throughout their genomes.

Table 13. Chromosome summary from SyMap with number of bison scaffolds placed on each domestic cattle chromosome.

Chromosome	Scaffolds placed
1	30
2	15
3	29
4	22
5	28
6	19
7	17
8	20
9	14
10	21
11	16
12	13
13	14
14	13
15	20
16	12
17	12
18	14
19	7
20	12
21	11
22	9
23	12
24	12
25	13
26	6
27	9
28	8
29	12
X	7
Total scaffolds placed	447

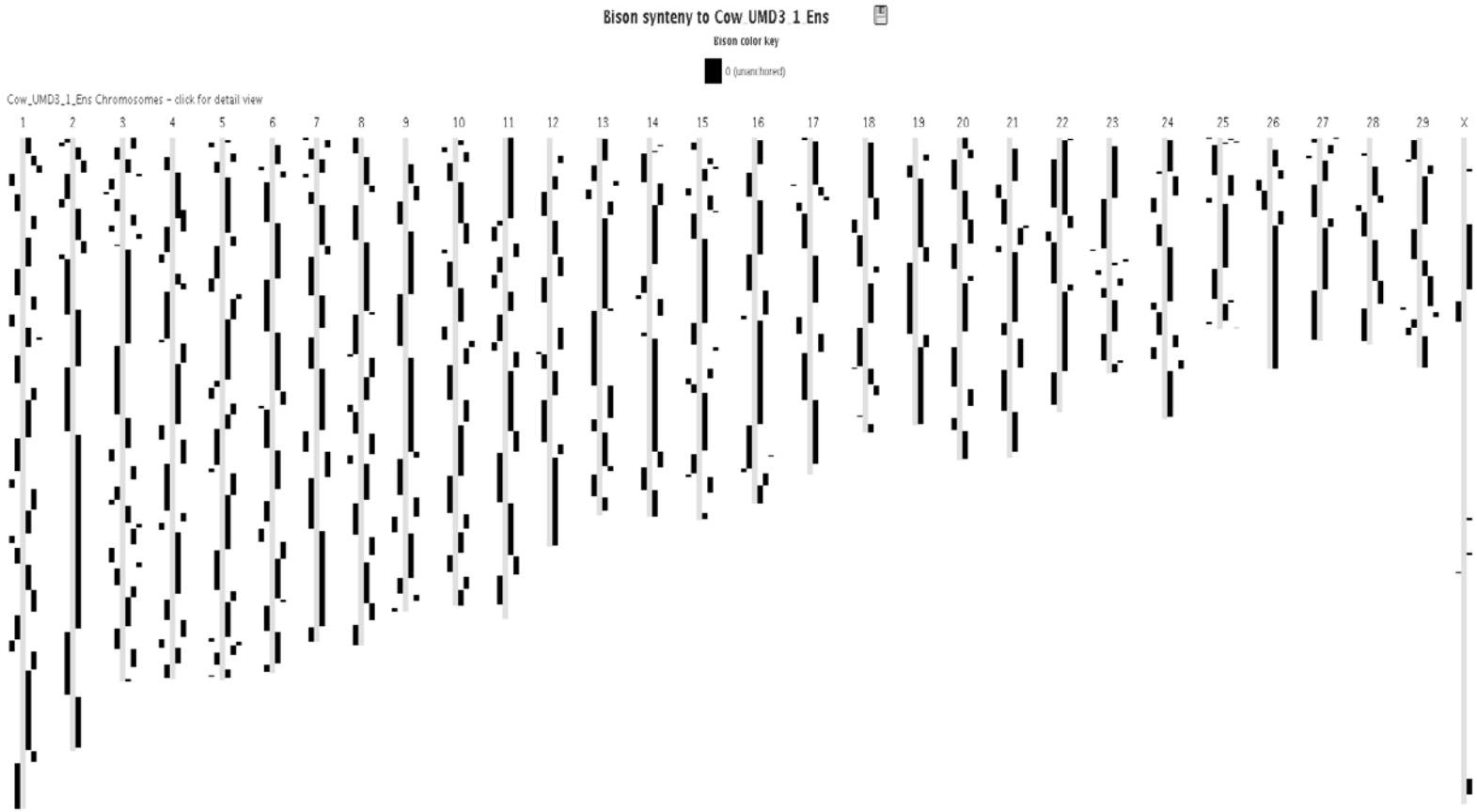


Figure 7. Bison synteny to domestic cattle UMD3.1.76. Black anchors are those scaffolds that were found to have synteny with domestic cattle.

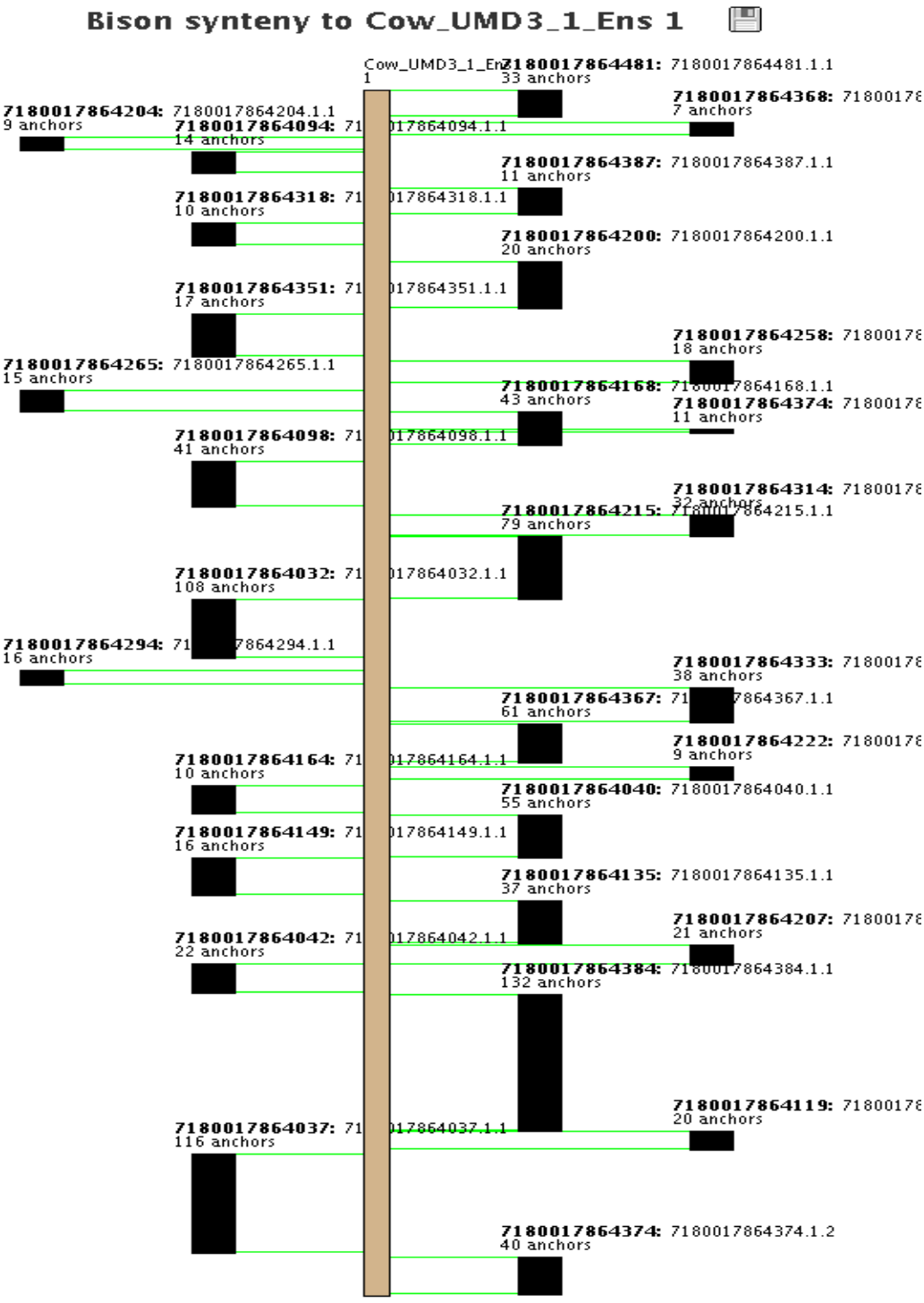


Figure 8. Bison scaffolds anchored to chromosome 1 of domestic cattle UMD3.1.76.

Discussion

With the completion of the 2.82-Gb *de novo* reference assembly of the American bison genome, bison genetic research has now advanced into the genomic technology era. The first genomic reference sequence of American Bison provides a deeper genomic evaluation for bison that currently used technologies cannot offer. With the annotation of the bison *de novo* reference genome we were able to identify a total of 26,001 genes and pseudogenes with 20,782 genes being protein coding genes. The function of these genes were also identified, which increased prior gene information greatly in bison.

The bison reference also provided a way to detect new genetic variants, including SNVs and INDELs, at the genomic level after being aligned to the domestic cattle reference. With over approximately 30,000,000 new variants (both SNVs and INDELs combined) found between bison and domestic cattle we have vastly expanded the number of variants that define the genomic differences between bison and domestic cattle. These identified genomic variations between bison and domestic cattle were then annotated to determine their associated genes and respective functions. In total, 24,497 genes were annotated from these variants, with the majority of the annotated genes being protein coding genes. We were able to identify 48 enriched gene ontology categories for biological pathways that produced with a False Discovery Rate P-value ≤ 0.05 comprising of 7,332 Ensembl Genes in the DAVID *Bos taurus* database. Using the bison reference genome and aligning it to the domestic cattle reference we were able to detect genetic variants and then annotate them to determine their gene function and

biological pathways they affect. By using whole genome sequencing technologies we have provided a ground-breaking analysis of the genomic differences that exist between bison and domestic cattle and the biological affects they could have in bison.

We were also able to identify those genomic components that are similar between the bison and domestic cattle reference that could be due to the ancestral influence of a shared common ancestor. This shared genomic information allowed us to use the domestic cattle reference to provide chromosomal assignments of the bison reference scaffolds. With no current chromosome map available for bison, we were able to utilize the bison scaffolds to provide location of genes on bison chromosomes without having to do additional chromosome mapping. We were able to anchor these genes to "pseudo-chromosomes" for bison using synteny blocks between the bison scaffolds and the domestic cattle chromosomes. These "pseudo-chromosomes" provide chromosome location for bison genes that may not have had prior information on. However, a future chromosome map for bison would provide a more thorough and precise mapping of bison genes onto chromosomes without using the domestic cattle as a reference.

With further whole genome sequencing of bison samples, a similar approach was used in the following chapter to align resequenced bison to the bison reference sequence to identify genomic variants between modern and historical bison. This multi-way comparison of bison genome sequences determined genomic differences between each bison population and the reference genome, as well as comparing these differences across populations. These variants can be used to determine evolutionary differences between modern and historical bison and offers the first comparative genomic analysis

of bison. These same bison sequences were also aligned to the domestic cattle reference sequence in the same fashion to further detect variants between bison and domestic cattle. With more bison sequences in the future we can validate these variants to be able to provide a genomic data set with known variants between bison and domestic cattle to be used for bison conservation management.

CHAPTER III

4 WAY GENOMICS COMPARISON OF NORTH AMERICAN BISON AND DOMESTIC CATTLE

Introduction

Bison bison *Classification*

The restoration of North American bison is considered as one of the first conservation success stories and is seen as a model of natural resource conservation (Ward 2000). To date there are approximately 500,000 bison in both private (raised as livestock) and conservation herds (Boyd 2003). Nearly all modern plains bison are descendants of the 76-84 bison that were used to establish the 5 private bison herds that aided in the recovery of American bison in the 1800's, along with the wild bison population in Yellowstone National Park (Garretson 1938; Meagher 1973; Coder 1975).

From a taxonomic point of view, North American bison are subdivided into two sub-species based on physical appearance and coat characteristics, the wood buffalo (*Bison bison athabascae*) and plains bison (*Bison bison bison*; Hall, 1981; McDonald 1981; Meagher 1986). Plains bison ranged historically across much of the United States and southwestern Canada, while wood buffalo occurred in north-western Canada; however the ranges of plains bison and wood buffalo were capable of overlapping (Potter *et al.* 2010). Subspecies were primarily assigned based on morphology (such as skulls, horns, and body proportions, size and hair patterns), but there is not a consensus as to if these designations are valid; as previous genetic studies have not supported the

distinction between plains and wood bison (McDonald 1981; Geist 1991; Cronin *et al*. 2013).

American plains bison were once grouped into two subspecies, northern plains bison (*Bison bison montanae*) and southern plains bison (*Bison bison bison*). These groupings were based on similar guidelines of wood buffalo and plains bison, by comparing physical appearance, horn, and coat characteristics depicted in pre-1900s illustrations (Krumbiegel and Sehm 1989). Hornaday (1886) also noticed similar differences in coat characteristics and attributed this to geographical and climate influences. If these subspecies classifications were in fact valid, since the 1900s they have crossbred freely and possibly eliminated these regional phenotypic differences between wood buffalo and plains bison, as well as southern and northern plains bison (Coder 1975; Dary 1989; McHugh 1972).

Unfortunately, the American colonization of North America in the late 1800s resulted in the almost complete elimination of the American bison (both wood buffalo and plains bison) and lead to the subsequent population bottleneck, reducing the population size by over 99.9% in less than 100 years (Coder 1975; Dary 1989). Wood buffalo numbers were down as low as 300 animals, relatives of these surviving animals are now found in the area belonging to Wood Buffalo National Park (Banfield and Novakowski 1960). Estimations of remaining plains bison ranged between a minimum of a few hundred individuals found in only 6 captive populations (Halbert 2003).

Charles Goodnight is noted as one of the five private ranchers who helped to protect American bison from extinction around the 1880's when hunters were

slaughtering all the adult bison, by starting his own small herd from 5 wild-caught orphaned calves (Coder 1975; Dary 1989). Goodnight's wild calves represented the last remaining examples of the southern extension of plains bison. The last remaining 36 bison would be used to create the Texas State Bison Herd (TSBH) after being relocated to Caprock Canyons State Park in 1997 (Swepston 2001). Therefore, the only modern remnant of the southern plains bison is believed to be found in the Texas State Bison Herd (Halbert 2003).

The collection of bison and establishment of the National Zoological Park was overseen by William T. Hornaday, who also realized the importance of saving the North American bison from extinction (Halbert 2003). Hornaday also collected bison hide, skull and skeletons that were archived at the Smithsonian Institute's Natural Museum of Natural History in order to preserve the legacy of bison for future generations (Hornaday 1886).

Since the early 1900's, the Canadian government has aided in the wood buffalo recovery by protecting the wild populations Wood Buffalo National Park from hunters (Halbert 2003). Wood Buffalo National Park was estimated to contain 1,500-2,000 bison by 1922, but despite this steady population increase and with many objections from Canadian scientists, approximately 6,600 plains bison were moved into the herd from 1925-1928 (Banfield and Novakowski 1960; Roe 1970). This would lead to the mixed breeding of wood buffalo and plains bison at Wood Buffalo National Park (van Camp 1989; Geist 1991), making it difficult to distinguish these hybrids from non-hybridized wood buffalo and plains bison. A pure sub-population of wood buffalo

(Banfield and Novakowski 1960) was believed to have been used to establish populations at Mackenzie Bison Sanctuary and Elk Island National Park in Canada (Geist 1991).

Microsatellite studies have found allele frequency differences between some herds of wood bison and plains bison, but all current wood bison populations have been shown to contain genetic material from plains bison (Cronin *et al*. 2013). Douglas *et al.* (2011) examined the complete mitochondrial DNA sequences of wood buffalo and plains bison, and found that the two wood bison haplotypes did not form their own clade; instead they were inter-mixed with the other 16 bison haplotypes. Cronin *et al.* (2013) concluded that the subspecies ranking of plains and wood bison was not supported by phylogenetic distinction and could be considered a northwestern (geographic) subpopulation of North American bison, fueling the debate that wood buffalo and plains bison should not be genetically distinct subspecies (Douglas *et al.* 2011).

Yellowstone National Park, founded in 1872, was the world's first National Park (Halbert 2003). However, poaching in Yellowstone National Park was widespread and President Cleveland enacted the Act to Protect the Birds and Animals in Yellowstone National Park and to Punish Crimes in Said Park and For Other Purposes, to punish those that committed wildlife related crimes in the park (Dilsayer 1994; Freese *et al.* 2007). By 1902 there were only 22 remaining wild bison in Yellowstone National Park (Garretson 1938; Meagher 1973; Halbert 2003). In that year, President Roosevelt appointed Charles "Buffalo" Jones game warden to help preserve the wild bison in

Yellowstone National Park. He played an integral part in supplementing additional outside animals from the Pablo-Allard (18 cows) and Charles Goodnight (3 bulls but one died) herds into the Yellowstone herd (Garretson 1938; Coder 1975; Halbert 2003). The supplemented bison were confined to paddocks and were managed as a captive herd and once numbers increased in 1915 they were released into the park and able to interact with the "wild" bison (Meagher 1973). The bison population at Yellowstone National Park is one of the most thoroughly studied and most well-known of the public bison herds in North America. Ward *et al.* (1999) found no evidence of domestic cattle mitochondrial DNA and 2 distinct haplotypes at Yellowstone National Park, as well as no detection of nuclear introgression of domestic cattle (Ward *et al.* 2000; Halbert 2003).

Bison Introgression with Domestic Cattle

The general consensus of the *Bison-Bos* genera split is that they once represented a single monophyletic clade believed to have derived from a common ancestor 0.5-2 million years ago in Eurasia (McDonald 1981). There is some disagreement over phylogenetic relationship among cattle and bison and most agree that the *Bison* genus should be included in the *Bos* genus (Simpson 1961; van Gelder 1977). The *Bison* genus is represented by two extant species, *Bison bison* (North American bison) and *Bison bonasus* (European bison; Both extant bison species can produce viable offspring with not only domestic cattle (*Bos taurus*) but other members of the *Bos* genus (Meagher 1986). Female progeny are fertile, while loss of fertility of male hybrid offspring can be restored with repeated back-crossings (Ward 2000; Verkaar *et al.* 2003).

It was well-known that the 5 cattlemen, who had helped with the recovery of bison, also had bison for the purpose of creating hybrids with domestic cattle to produce a better meat source, or beefalos (Coder 1975). Hybridization, whether forced or spontaneous between bison and domestic cattle as well as other bovine species, might compromise the genetic integrity of bison (Verkaar *et al.* 2003). The association with domestic cattle mitochondrial DNA and reduced body size in bison was able to show that genetic introgression from domestic cattle does have an effect on bison (Derr *et al.* 2012).

This hybridization of domestic cattle into bison presents challenges in the management and conservation of the American bison today, because most advanced generation backcrosses are morphologically indistinguishable from purebred bison (Douglas *et al.* 2011). Freese *et al*. (2007) report that we can at best say that less than 1.5% of the 500,000 plains bison in existence today can be considered as likely free of domestic cattle introgression. Current research has found that modern bison herds do possess both mitochondrial (Polziehn *et al.* 1995; Ward *et al.* 1999) and nuclear domestic cattle DNA (Ward 2000; Halbert 2003; Halbert *et al.* 2005). While these technologies are useful for detecting introgression within herds (e.g., >100 bison), they do not provide the needed resolution to detect cattle introgression in individual bison at the genomic level.

Advancing Bison Management

Whole genome sequencing has been used to identify levels of introgression of Asian haplotypes in European breeds of pigs (Bosse *et al.* 2014). Chinese pigs were

known for having great mothering characteristics, superior meat quality, strong resistance to diseases, better adaptation to living in sties, and producing larger litters (>15 live born piglets(young); Bosse *et al.* 2014). In the early eighteenth and nineteenth centuries, Chinese breeds were brought to Europe to help improve commercial traits such as meat quality, development and fertility were chosen to be placed into European breeds by breeders in European breeds by European breeders (Bosse *et al.* 2014). Using whole-genome sequencing data, levels of introgression of Asian haplotypes in European breeds were identified by associating regions where genes controlling certain associated production phenotypes were associated with Asian haplotypes; this is an example of purposefully adding genetics of different breeds to try and produce a more efficient breed for livestock purposes (Bosse *et al.* 2014).

Using a similar approach with whole genome sequencing technologies, we assessed the genetic variants among bison and between bison and domestic cattle. The cattle genome UMD3.1 (Ensembl GCA_000003055.3), recently completed reference bison genome (Bison_UMD1.0), two historic bison samples that predate introgression from the bison collection at the Smithsonian Institution, four wood buffalo samples from Elk Island National Park (EIW), four bison samples from Caprock Canyons State Park (CCSP) and 4 Yellowstone National Park (YNP) bison samples were used to identify genomic variants, including SNPs and INDELs and, between domestic cattle and bison. Variants between each population were compared to variants in historic bison to provide a new list of genomic variants between bison and domestic cattle. These variants were then annotated to determine the effects on gene structure, and what variants could be

controlling genes that affect phenotypic traits or regulation in bison. Biological processes, such as functions in regulation of processes, transcription, and development, enriched for these variants were analyzed and compared between bison and domestic cattle. This 4 way comparison has identified genes and their respective functions from variants detected for bison and domestic cattle that can be used in future research to better understand how these genes function differently in bison.

Whole-genome sequencing provides the next step in advancing bison management and conservation. With the completion of the *de novo* plains bison reference genome sequence we compared the same historic and modern bison sequences and identified genomic variants and their associated functional genes. The two historic bison samples were used to determine conserved genetics between historic and current bison sequences. Bison sequences from wood buffalo samples EIW, CCSP and YNP were compared for variant identification to distinguish unique genes for each population. This study provides an outline of the genetic architecture of bison that existed before the population bottleneck and a more in-depth genomic analysis that will be used for bison conservation management of populations of bison. The utilization of genomic technology with this iconic species allows insight into the genetic history, taxonomy, and inheritance of important genetic traits in bison that have allowed them to thrive over the years.

Materials and Methods

Collection of Historic DNA Samples/Isolation of DNA of Historic Samples

Historic bison samples were from the bison collection from the Smithsonian Museum of Natural History Archives and met our criteria for candidates. These criteria consisted of the sample being well documented regarding location, date, and collector and it must have been collected before extensive hybridization occurred between bison and domestic cattle. This is important to ensure that we can define the bison genome that existed before introgression and the population bottleneck. This allowed us to provide genome sequences of bison that would be without known introgression in order to provide a foundation of bison genomics that would be the standard to evaluate introgression of domestic cattle genetics into bison.

Two samples were chosen, one female skull sample 6 (Smithsonian Institute ID 015696) that was collected November 3, 1886 by Hornaday, in Dawson County, Montana. The second sample was skull sample 9 (Smithsonian Institute ID 002007) and was collected in August 1856 by Hayden in the area that would become the Hayden Valley in central Yellowstone National Park. To ensure that contamination of these historical samples with modern bison DNA, all historical samples were handled outside of our lab, which deals with genetic testing of modern bison samples. Due to the age, degradation, and importance of extracting good quality DNA of these samples, DNA was extracted at the North Texas – Health Science Center DNA Forensics lab under the direction of Dr. Bruce Budowle, which deals with human forensics samples that are often highly degraded using their extraction protocol.

Bone Preparation

The outer surfaces of the bone fragments were cleaned by immersing them in 50% commercial bleach (3% NaOCl) in a 50-mL conical tube for15 min. Next, the bones were briefly washed with nuclease-free water (4–5 washes). The bone fragments were immersed briefly in 95–100% ethanol and air dried overnight in a sterile hood. The bone fragments were pulverized using a 6750 Freezer/Mill, using a protocol of a 10-min rechill followed by 5 min of grind time at 15 impacts per second (SPEX SamplePrep L.L.C., Metuchen, NJ, USA).

Hi-Flow® Silica-Column Extraction

The Hi-Flow columns (purchased from Generon Ltd.), were constructed on the 20 mL capacity Proteus™ (AbD Serotec, Raleigh, NC) protein purification column platform (designed to be seated in a 50 mL conical tube during use) and contain a glass fiber filter. The chemistry for the Hi-Flow protocol is similar to that with the QIAquick® (Qiagen, Valencia, CA) silica gel columns (as modified from Yang et al. [8]). Bone demineralization was carried out by mixing approximately 0.5 g bone powder with 3 mL digestion buffer (0.5 M EDTA pH 8.0; Invitrogen Corporation, Carlsbad, CA), 1% sodium N-lauroylsarcosinate (Sigma-Aldrich Corp., St. Louis, MO) and 200 µL of proteinase K (Roche Applied Science, Indianapolis, IN) (20 mg/mL), followed by incubation in a hybridization oven at 56°C under constant agitation overnight. After demineralization, the bone powder was pelleted via centrifugation at 2545 x g for 5 min. The supernatant was transferred to a sterile conical tube and mixed with five volumes of binding PB buffer (Qiagen). This mixture was vortexed thoroughly, transferred to a Hi-

Flow DNA Purification Spin Column, and centrifuged at 2545 x g for 10 min. After discarding the eluate, the column was washed with 15 mL PE buffer (Qiagen), centrifuged at 2545 x g for 5 min and washing repeated for a total of three washes. The empty column was centrifuged at 2545 x g for 5 min to remove residual ethanol from the PE buffer. The column was transferred to a sterile collection tube, and the DNA was eluted with 100 µL elution buffer (EB, Qiagen). Three elutions were performed for each sample for a total recovered volume of approximately 300 µL for each bone. Each elution was transferred to a separate, sterile1.5mL microfuge tube. The DNA extracts were stored at 4°C and −20°C for short- and long-term storage, respectively. And DNA quality was considered before library preparation on an Agilent Tape Station following manufacturer's protocol (Agilent Technologies, Santa Clara, CA).

Sample Collection and DNA Extraction for EIW, CCSP, and YNP Bison

Four wood buffalo from Elk Island National Park (EIW) in Alberta, Canada were chosen for genomic sequencing due to reports that they represent a pure wood buffalo population (Geist 1991). Four animals were also selected to represent what is believed to be the last remaining population of southern plains bison from the Texas Parks and Wildlife managed Caprock Canyons State Park (CCSP) outside of Quitaque, Texas. Lastly, four bison samples were chosen to represent Yellowstone National Park (YNP) for genomic sequencing since the bison herd at YNP represents what is documented as one of the only bison herds not to have domestic cattle introgression detected. Tail hair samples from these 12 samples were extracted by MasterPure DNA Purification Kit (Epicentre Biotechnologies, Inc., Madison, WI). And DNA quality was considered

before library preparation by Quant-iT PicoGreen dsDNA Assay Kit. (LifeTechnologies, California).

Whole-Genome Re-sequencing

Illumina paired-end libraries were prepared for sequencing on the Illumina HiSeq 2000^{TM} Next-Gen from the above extracted DNA for whole genome resequencing using the Nextera DNA Sample Preparation Kit (Illumina, San Diego, CA). For each of the 4 samples from EIW and CCSP the genomic libraries were indexed with adapters and four samples were run together on 2 HiSeq lanes samples using the 2x100 normal mode. This generated approximately 5X coverage for each sample. The historic samples were not combined due to lower quality DNA and libraries were prepared using the NEXTflex Illumina ChiP-Seq Library Prep Kit by Bioo Scientific (Bioo Scientific Corporation, Austin, TX) protocol and ran on one lane with the normal mode High Output 2x100 mode (Illumina, San Diego, CA). Illumina TruSeq Nano libraries for the 4 samples from YNP were prepared using the Illumina TruSeq Nano DNA Sample Preparation Kit (Illumina, San Diego, CA), and ran on 4 separate lanes on 2x100 mode.

Samples were also blasted using the NCBI blast nucleotide database command line option to detect any foreign sequences (Camacho *et al.* 2009; Altschul *et al.* 1997; Altshcul *et al.* 1990). These sequences were removed during the filtering process described below.

Sequence Alignment

Prior to aligning the historic sequences and YNP samples to the bison reference sequences, sequences were trimmed using FASTQ-MCF, filtering out both bases with a

quality score less than 20 from each individual read and reads with a remaining sequence length of less than 70 bases (Aronesty 2011). Whole Systems Genomics Initiative (WSGI) provided the computational resources and systems administration support for the WSGI HPC Cluster used for these analyses. These filtered paired-end sequences along with wood buffalo and Caprock Canyons State Parks sequences were individually aligned to the reference bison scaffolds and domestic cattle (UMD3.1) reference sequence using Burrows-Wheeler Alignment 0.6.2 (BWA-MEM; Li 2013) using the default settings. The resulting BAM (binary short DNA sequence read alignment; Li *et al.* 2009) files were combined using the merge option of the Sequence Alignment/Map (SAM)tools 0.1.18 software package (Li *et al.* 2009). Read group information was added using the AddOrRelpaceReadGroups option of PicardTools 1.7.1 (https://github.com/broadinstitute/picard/releases/tag/1.128). Then Genome Analysis Toolkit 3.1.1 (GATK; McKenna *et al.* 2010) option RealignerTargetCreator was used to realign and account for INDEL shifted coordinates to create a realigned and sorted BAM file of read alignments to bison scaffolds and domestic cattle (UMD3.1) for each sample. Finally the SAMtools view and flagstat options (Li *et al.* 2009) were used to obtain statistics of the alignments of individual samples with bison and domestic cattle reference genomes.

Identification of Genetic Variants and Analysis

Genetic variants, SNVs and INDELs, were identified against both the bison and domestic cattle references for each aligned sample and were filtered according to the GATK Best Practices recommendations (DePristo *et al.*, 2011; Van der Auwera *et al.*,

2013). The resultant variants were placed into variant call formatted (VCF) files. VCFtools merge and SAMtools bcftools merge (Danecek *et al.* 2011) options were used to combine the individual SNV or INDEL VCF files to the corresponding populations, to give a VCF file of all samples SNVs or INDELs into one table (i.e. a table for all samples for EIW samples and their SNVs were combined into one table) for comparison to the bison and then domestic cattle reference sequence. VCFtools 0.1.11 vcf-stats (Danecek *et al.* 2011) option was used to determine basic statistics and counts of the SNVs and INDELs for each population, and the historical samples.

In order to annotate the identified variants for the bison populations to the bison reference, the SyMap pseudo-chromosomes were used to change the scaffold IDs in the combined bison population VCF to actual chromosome numbers based on position since a bison reference is not available in SnpEff to annotate variants. The combined VCF variants were then annotated using SnpEff 4.1 software (Cingolani *et al.* 2012) against the UMD3.1.76 reference from Ensembl. SnpEff 4.1 was also used to annotate the identified variants to the domestic cattle reference against UMD3.1.76 reference. The annotated variants were then analyzed using the DAVID Functional Annotation Tool (FAT; Huang *et al.* 2009; http://david.abcc.ncifcrf.gov/home.jsp) in order to identify enriched biological pathways for the variants called to the domestic cattle and bison references.

Phylogenetic Analysis

SNPhylo version 20140701 (Lee *et al.* 2014) was used to generate a phylogenetic tree using the combined VCF file to domestic cattle (UMD3.1). The VCF file to UMD

3.1 was chosen for this analysis and not the combined VCF file UMD1.0 so Templeton would be included in the analysis.

Results

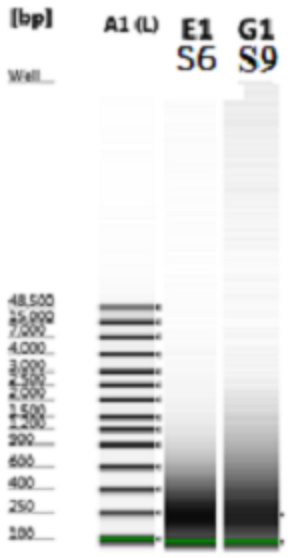
Information for Sequenced Samples

For the 15 samples used for sequencing, both re-sequencing and for the *de novo* assembly, Table 14 offers the Sample ID, Alternate ID, Location, Sex, Notes, and what tissue the DNA was extracted from. Of the 15 samples, the majority of the samples used were males, with only 5 samples being female, and one sex was unknown. All of the sequences will be deposited in the SRA database at NCBI following the guidelines for submitting sequences.

Table 14. Sample information for the 15 bison samples used for sequencing analysis.

Sample ID	Alternate ID	Location	Sex	Notes	DNA Extracted From
26-1525	2013001525	EIW	Male		Hair Follicles
95-1573	2013001573	EIW	Male		Hair Follicles
151-1607	2013001607	EIW	Male		Hair Follicles
233-1676	2013001676	EIW	Male		Hair Follicles
50-5792	2010005792	CCSP	Male		Hair Follicles
61-5793	2010005793	CCSP	Male		Hair Follicles
48-5795	2010005795	CCSP	Male		Hair Follicles
68-5784	2010005784	CCSP	Male		Hair Follicles
YNP1856	13927	YNP	Female		Hair Follicles
YNP1861	13932	YNP	Female		Hair Follicles
2009005885	16-09	YNP	Female		Hair Follicles
2009005899	08-09	YNP	Female		Hair Follicles
Templeton	2011002044	YNP	Male		Whole Blood
S6	015696	Dawson County, MT	Female	Historical Sample collected Nov 3, 1886 by W. T. Hornaday	Skull bone pieces
S9	002007	Yellowstone	Unknown	Historical Sample collected Aug 1856 by F. V. Hayden	Nasal Cavity pieces

To ensure that the extracted genomic DNA for the historic samples was of good quality to be used for preparing whole genome sequencing libraries, Agilent Tape Station High Sensitivity Tape was used to analyze the quality of the genomic DNA (Figure 9). Historical samples, S6 (lane E1) and S9 (lane G1) were found to have not only the proper concentrations needed for ChiP-Seq library preparation, but they also had good quality. Figure 9 also offers the amount of concentration for the standard used in lane A1 of the Tape Station gel at each size fragment, and a total of 61.78 ng was used in 1 µL of the standard.



Size [bp]	Calibrated Conc. [ng/µl]
100	8.5
250	4.63
400	4.57
600	4.42
900	4.46
1200	4.23
1500	4.29
2000	3.99
2500	3.88
3000	4.05
4000	3.87
7000	3.83
15000	3.82
48500	3.24
Total Conc.	61.78 in 1µL

Figure 9. Agilent Tape Station results for assessing quality of genomic DNA of historical samples S6 and S9.

Samples were also blasted to the nucleotide database at NCBI, using the command line option on the Texas A&M Institute for Genome Sciences and Society server. Table 15 offers examples of the main hits for the historical sample S6, as well as hits for other foreign sequences. Some human sequences were found have blast hits, as well as pig, and a parasite. The majority of the top hits belonged to pseudogenes in *Bos taurus*. Since the historic samples were expected to have more foreign sequences due to the age and deterioration of the samples, it is noteworthy that the majority of the blast hits were found to be most similar to *Bos taurus* and few hits were bacterial or human.

Table 15. Partial blast report for historical sample S6.

Locus	Hits	Species	Notes
JX848345.1	25,471	*Bos taurus*	pseudogenes
JQ711177.1	10,499	*Bos taurus*	growth hormone receptor gene
LL714604.1	1	*Elaeophora elaphi*	nematode parasite
BC079477.1	1	*Homo sapiens*	ribosomal protein L37, mRNA
NG_033880.1	1	*Homo sapiens*	protein tyrosine phosphatase
NG_021367.1	1	*Homo sapiens*	SH3-domain kinase binding protein 1
CU467051.7	1	*Sus scrofa*	pig DNA sequence from clone

Variant Identification to Templeton

Wood Buffalo (EIW)

The mem alignment option of BWA was used to align the raw sequences of the 4 wood buffalo (EIW) samples using Illumina sequencing technology to the bison reference genome sequence (UMD1.0; Templeton). The SAMtools options, view and

flagstat, were used to obtain statistics of the 4 wood buffalo Illumina paired-end reads mapped to the bison reference sequence. Two of the four samples (sample IDs 95-1573 and 151-1607) were found to have fewer total reads to be aligned to the bison reference sequence, due to lower quality of sequences (Table 16). Even though the reads for each sample ranged from 12,968,260 (95-1593) to 75,729,836 (sample 233-1676), there was only an average of 41,313,316 reads to compare to bison reference sequence.

Table 16. Samtools flagstat statistics of the 4 EIW bison raw sequences mapped to the bison reference sequence UMD1.0 (Templeton; reads are in base pairs).

Statistic	26-1525	95-1573	151-1607	233-1676
in total (QC-passed reads + QC-failed reads)	58,711,530	12,968,260	17,859,638	75,729,836
duplicates	0	0	0	0
Mapped	41,146,430	9,748,583	14,277,587	42,299,806
(% mapped)	(70.08%)	(75.17%)	(79.94%)	(55.86%)
paired in sequencing	58,711,530	12,968,260	17,859,638	75,729,836
read1	29,355,765	6,484,130	8,929,819	37,864,918
read2	29,355,765	6,484,130	8,929,819	37,864,918
properly paired	32,641,705	7,378,686	11,124,593	34,112,748
with itself and mate mapped	36,899,718	8,190,378	12,380,381	38,841,851
singletons	4,246,712	1,558,205	1,897,206	3,457,955
with mate mapped to a different chr	3,549,738	820,399	1,253,147	3,069,957
with mate mapped to a different chr (mapQ>=5)	1,748,202	395,165	616,363	1,405,448

SNVs and INDELs were called and identified separately using GATK against the bison reference genome. The same two samples that were found to have fewer reads for alignment (sample IDs 95-1573 and 151-1607) were found to have fewer variants, both

SNVs and INDELs, which is expected since fewer reads were able to be used for analysis. Table 17 offers a summary for each individual sample and their variants detected, where the most SNVs detected were found for wood buffalo 26-1525 with 1,142,446 SNVs. The most detected INDELs were found for wood buffalo sample 233-1676 with 7,513 INDELs detected. Most of the variants detected were homozygous variant alleles. Since these samples were compared to a sub-species we would expect to see some alleles from the bison reference genome, which can be found in the heterozygous variants, where one variant was the reference (bison) allele), and not a significant amount of variants.

Unique alleles are those variants that were detected in only one individual and not the other three when statistics were ran on the combined VCF population. There was a considerable amount of unique alleles that were detected for each individual wood buffalo sample. This could be due to sequencing errors or lower quality variants being called, which were removed for the rest of the analysis.

Table 17. Individual variant summary statistics of 4 wood buffalo found from comparing sequences to Templeton.

	26-1525		95-1573		151-1607		233-1676	
	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs
Homozygous Variant alleles	834,608	4,412	103,913	625	216,162	861	766,704	6,046
Heterozygous (one Reference one variant)	307,466	1,179	33,286	190	60,180	306	257,234	1,435
Variant Count	1,142,446	5,626	137,317	825	276,476	1,181	1,024,294	7,513
Reference Alleles	307,466	1,179	33,286	190	60,180	306	257,234	1,435
Unique Alleles	877,084	3,920	66,683	80	155,558	221	779,746	5,888
Heterozygous Variant Alleles	372	35	118	10	134	14	356	32

Using SAMtools bcftools merge option the wood buffalo variant files were combined to take into account these variants for the population as a whole for each genetic variant type. Analysis of summary statistics was completed using Samtools vcf-stats and confirmed by Samtools bcftools stats and SnpEff (Tables 18 and 19) for the variants detected between wood buffalo and Templeton. The main substitutions for the SNVs found for wood buffalo samples were transitions of T>C and C>T and the least common base substitution was found to be a T>A transversion (Table 18). The transition (Ts) to transversion (Tv) ratio was 2.07. For whole genome studies a Ts/Tv ration is expected to be between 2-2.1, suggesting that few false positives generated by random sequencing errors were within the sequence; our Ts/Tv ratio of 2.06 for EIW population SNVs confirms that we were able to properly detect variants with few false positives between wood buffalo and Templeton (Li 2011). Individual Ts/Tv ratio for the 4 EIW bison is shown in Table 18, and the ratios for the 2 samples that had lower reads to start with had lower ratios than the other 2 samples. Although these were lower they were still able to detect variants without false positives and any low quality variants were properly excluded from further analysis.

Table 18. Base substitution counts for SNVs identified for 4 wood buffalo from alignment to Templeton.

	A	C	G	T
A	0	95,525	368,230	83,283
C	94,439	0	85,853	370,687
G	368,028	85,974	0	96,450
T	81,140	371,309	91,700	0

Table 19. Transition and transversion counts and ratio for each EIW sample and population total found against Templeton.

Sample	26-1525	95-1573	151-1607	233-1676	Total
Transitions	1,340,885	155,024	325,333	1,208,913	1,478,254
Transversions	636,541	86,324	167,439	582,441	714,364
Ts/Tv	2.11	1.80	1.94	2.08	2.07

The corresponding quality score for each variant called can be found in Figure 10. In total there were 2,189,369 SNPs and 11,931 INDELs, with 6,408 insertions and 5,593 deletions, detected between wood buffalo and the bison reference sequence. Only 21,683 of the 2,189,369 (1.0%) SNPs were found to be informative SNPs, meaning that these SNPs were found in all 4 wood buffalo samples and only 496 informative INDELs (Table 20). Overall, there were 2,204,619 variants detected between wood buffalo and the bison reference sequence.

Table 20. Number of common variants found between the 4 EIW samples.

Shared between	SNPs	INDELs
4	21,683	496
1	1,879,071	10,109
3	37,500	400
2	251,115	926
Total	2,189,369	11,931

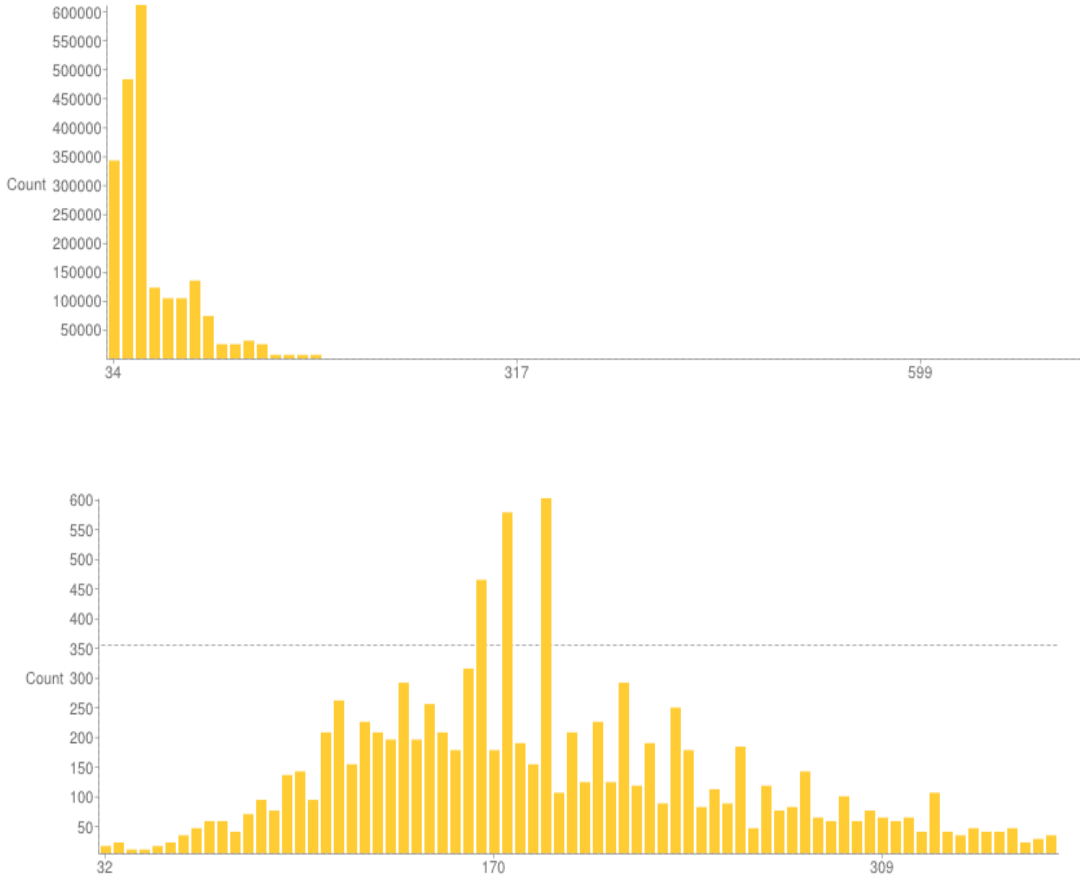


Figure 10. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for wood buffalo aligned to the reference bison.

Caprock Canyons State Park (CCSP)

Raw paired-end sequences of 4 CCSP bison were individually aligned to the bison reference sequence similar to that as stated above for wood buffalo. The analysis for variants will remain separate for SNVs and INDELs and then grouped together to

consider the variants within the population as a whole, the same as the results for wood buffalo.

The raw reads for comparison were more even for these 4 CCSP samples than they were for the wood buffalo comparison. Using the SAMtools option flagstat was used to obtain statistics of the 4 CCSP Illumina paired-end reads mapped to the bison reference sequence (Table 21). There was an average of 69,206,349 raw Illumina parried-end reads for the CCSP bison to be compared to the bison reference sequence.

Table 21. Samtools flagstat statistics of the 4 CCSP bison raw sequences mapped to Templeton (reads are in base pairs).

Statistic	48-5795	50-5792	61-5793	68-5784
in total (QC-passed reads + QC-failed reads)	74,724,288	71,988,836	66,683,988	63,428,284
duplicates	0	0	0	0
Mapped (% mapped)	52,144,062 (69.78%)	52,244,308 (72.57%)	44,460,951 (66.67%)	43,771,930 (69.01%)
paired in sequencing	74,724,288	71,988,836	66,683,988	63,428,284
read1	37,362,144	35,994,418	33,341,994	31,714,142
read2	37,362,144	35,994,418	33,341,994	31,714,142
properly paired	39,668,810	42,950,890	35,374,562	35,404,262
with itself and mate mapped	44,348,057	48,003,551	39,765,494	39,499,941
singletons	7,796,005	4,240,757	4,695,457	4,271,989
with mate mapped to a different chr	4,134,640	4,091,326	3,404,780	3,171,155
with mate mapped to a different chr (mapQ>=5)	1,848,354	1,948,418	1,582,185	1,518,451

Table 22 offers a summary for each individual sample and their variants detected, where the most SNVs detected were found for CCSP animal 50-5792, with 1,635,988 SNVs detected. The most detected INDELs were found for wood buffalo sample 48-

5795 with 11,217 INDELs detected. Similar to that found for wood buffalo, most of the variants detected were homozygous variant alleles. Since these samples were compared to the same species we would expect to see some alleles from the bison reference genome, which is shown through the heterozygous variants with one reference (bison) allele. The same trend that was seen for the 4 EIW bison and unique alleles were seen for the 4 CCSP samples. Each sample was found to have their own variants (unique to them) that were not seen for the other CCSP samples (Table 22).

Table 22. Individual variant summary statistics of 4 CCSP found from comparing sequences to Templeton.

	48-5795		50-5792		61-5793		68-5784	
	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs
Homozygous Variant alleles	1,108,254	9,277	1,241,895	8,711	972,251	5,629	1,050,686	6,390
Heterozygous (one Reference one variant)	334,769	1,887	393,615	1,648	297,128	1,276	263,463	1,097
Variant Count	1,443,490	11,217	1,635,988	10,399	1,269,794	6,948	1,314,555	7,533
Reference Alleles	334,769	1,887	393,615	1,648	297,128	1,276	263,463	1,097
Unique Alleles	658,318	8,017	791,906	7,209	539,821	4,036	587,341	4,844
Heterozygous Variant Alleles	467	53	478	40	415	43	406	46

Using SAMtools bcftools merge option the CCSP bison samples individual variant files were combined using SAMtools bcf-merge to take into account these variants for the population as a whole. The main substitution for the SNVs found for CCSP samples was a transition of C>T and the least common base substitution was

found to be a T>A transversion (Table 23). The transitions to transversions ratio for the CCSP bison were 2.12 and each individual Ts/Tv ratio can be found in Table 24. All of the CCSP Ts/Tv ratios were around 2.1 and indicates that variants were called with few false positives. Any low quality variants were removed from for further analysis.

Table 23. Base substitution counts for 4 CCSP bison sequences aligned to Templeton.

	A	C	G	T
A	0	165,125	651,395	144,620
C	165,891	0	149,977	666,287
G	659,784	148,560	0	168,185
T	142,790	655,595	159,528	0

Table 24. Transition and transversion counts and ratio for each CCSP sample and population total found against Templeton.

Sample	48-5795	50-5792	61-5793	68-5784	Total
Transitions	1,740,264	1,964,470	1,523,979	1,608,155	2,633,061
Transversions	811,947	913,891	718,481	757,492	1,244,676
Ts/Tv	2.14	2.15	2.12	2.12	2.12

The quality for each variant count for the CCSP sequenced population can be found in Figure 10. In total there were 3,872,780 SNPs and 28,333 INDELs, with 14,769 insertions and 13,683 deletions, detected between CCSP bison and the bison reference sequence. Only 103,125 of the 3,872,780 (2.66%) SNPs were found to be informative, or found to be in common in all 4 CCSP samples, and only 1,394 INDELs (Table 25). Overall, there were 3,906,189 variants detected between CCSP bison and Templeton.

Table 25. Number of common variants found between the CCSP bison.

Shared between	SNPs	INDELs
4	103,125	1,394
1	2,577,386	24,106
3	289,403	749
2	902,866	2,084
Total	3,872,780	28,333

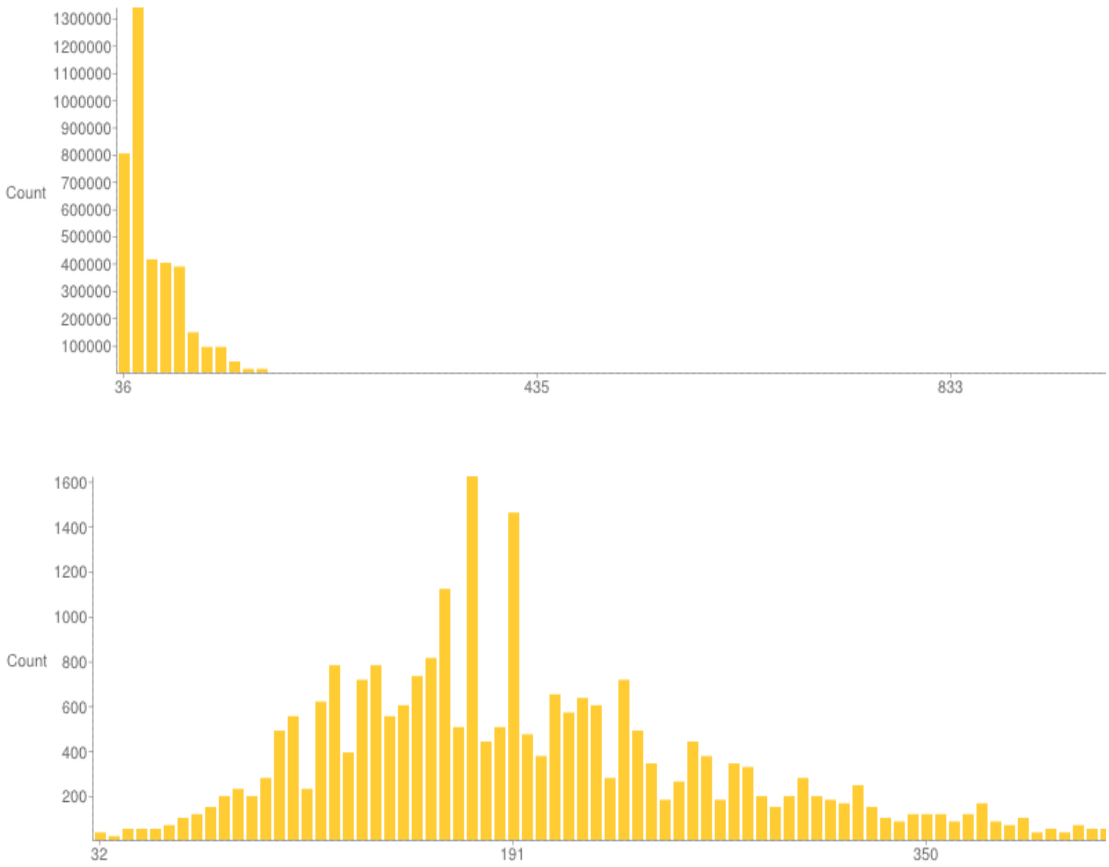


Figure 11. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for CCSP bison aligned to Templeton.

Yellowstone National Park (YNP)

The YNP samples were trimmed using FASTQ-MCF with a quality score less than 20 from each individual read and minimum remaining sequence length of 70 (Aronesty 2011), and then aligned individually to Templeton. The variant detection for the YNP samples would be expected to be lower than that of the other populations since these bison samples came from the same population as Templeton. This analysis shows how samples from the same population can still have genomic variants detected at a high occurrence.

The mem alignment option of BWA was used to align the filtered raw sequences of the YNP samples using Illumina sequencing technology to Templeton. Samtools option, view and flagstat were used to obtain statistics for reads used for the alignments for each individual sample (Table 26). An average of combined reads of 212,061,529 for the 4 YNP samples was used to detect genomic variants to Templeton and on average approximately 93% of the reads was properly mapped to Templeton.

Table 26. Samtools flagstat statistics of the 4 YNP bison raw sequences mapped to Templeton (reads are in base pairs).

Statistic	YNP1856	YNP1861	2009005885	2009005899
in total (QC-passed reads + QC-failed reads)	160,923,224	276,808,942	190,921,092	168,454,552
duplicates	0	0	0	0
mapped	158,441,204 (98.46%)	272,498,752 (98.44%)	187,474,083 (98.19%)	165,824,518 (98.44%)
paired in sequencing	160,923,224	276,808,942	190,921,092	168,454,552
read1	80,461,612	138,404,471	95,460,546	84,227,276
read2	80,461,612	138,404,471	95,460,546	84,227,276
properly paired	147,296,207	252,199,384	176,039,405	155,903,495
with itself and mate mapped	157,045,725	270,016,334	185,951,243	164,489,361
singletons	1,395,479	2,482,418	1,522,840	1,335,157
with mate mapped to a different chr	10,067,739	18,545,651	9,860,702	8,613,391
with mate mapped to a different chr (mapQ>=5)	4,435,469	8,045,289	4,597,635	3,975,717

Table 27. Individual variant summary statistics of 4 YNP bison found from comparing sequences to Templeton.

	YNP1856		YNP1861		2009005885		2009005899	
	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs
Homozygous Variant alleles	2,381,457	111,780	1,127,052		2,377,415	149,408	2,275,988	119,571
Heterozygous (one Reference one variant)	2,506,932	40,731	1,639,910		2,901,569	72,585	2,719,621	52,739
Variant Count	4,889,251	152,573	2,768,161		5,280,175	222,132	4,996,519	172,381
Reference Alleles	2,506,932	40,731	1,639,910		2,901,569	72,585	2,719,621	52,739
Unique Alleles	1,162,555	79,568	866,529		1,683,648	156,426	1,449,246	106,675
Heterozygous Variant Alleles	862	62	1,199		1,191	139	910	71

After using GATK to detect SNVs and INDELs for the mapped reads, vcftools stats was used to evaluate individual statistics of the genomic variants detected. Most of the SNVs detected for each individual were heterozygous SNVs, for a reference and a variant allele (Table 27). The YNP samples are expected to have some alleles in

common with Templeton since they are from the same population. For each of the 4 YNP samples the reference alleles detected were approximately 51.3%, 59.2%, 55.0%, and 54.4% of the total alleles detected for YNP1856, YNP1861, 2009005885, and 2009005899, respectively. The variants detected shows how samples from the same population can still have over a million SNVs detected within an individual genome.

As was done for the other population samples above, SAMtools vcf-stats and SNPEff were used to analyze and obtain statistics for combined VCF file of genomic variants detected, after being combined by the merge option in bcftools. This enabled us to evaluate population statistics for the variants detected. The main base substitution between the 4 YNP samples and Templeton was found to be a transition from C>T with 1,652,008 detected (Table 28). The individual Ts/Tv ration can be found in Table 29, along with the total population transition and transversions found for the YNP samples. The combined Ts/Tv ration for the YNP samples was 2.21, which is similar to the Ts/Tv rations listed above for the other populations, and also suggests were able to detect SNPs without calling false SNPs.

Table 28. Base substitution counts for 4 YNP bison from alignment to Templeton.

	A	C	G	T
A	0	367,641	1,511,779	314,904
C	389,725	0	361,640	1,652,008
G	1,637,017	358,495	0	393,308
T	311,346	1,509,408	364,272	0

Table 29. Transition and transversion counts and ratio for each YNP sample and population total found against Templeton.

Sample	YNP1856	YNP1861	2009005885	2009005899	Total
Transitions	5,008,016	2,653,119	5,272,157	5,013,185	6,31,0212
Transversions	2,263,554	1,243,293	2,386,624	2,260,232	2,861,331
Ts/Tv	2.21	2.13	2.21	2.22	2.21

The quality scores for those variants identified, both SNPs and INDELs, can be found in Figure 12. Table 30 shows the amount of variants found to be in common between the YNP samples. A total of 7,995,395 SNPs were found for the YNP samples to Templeton. There were a total of 741,721 (8.1%) of the total SNPs found to be informative, or in common between all 4 YNP samples. YNP1861 was not included for the INDEL analysis, due to after multiple attempts of running GATK no INDELs were produced in the VCF file. Therefore for the remaining 3 YNP samples, YNP1856, 2009005885 and 2009005899, a total of 408,375 INDELs were detected between the YNP samples and Templeton, with 168,588 (51.1%) insertions and 161,471 (48.9%) deletions. Only 22,802 INDELs were found to be in common between (informative) the 3 YNP samples. In total there were 9,566,325 genomic variants identified between the YNP samples, but only 764,523 (8.0%) of the total variants were found to be in common in all 4 YNP samples.

Table 30. Number of common variants found between the 4YNP samples for SNPs and only 3 YNP samples for INDELs.

Shared Between	SNPs	INDELs
4	741,721	
1	3,834,283	292,466
3	1,969,047	22,802
2	2,612,899	93,107
Total	9,157,950	408,375

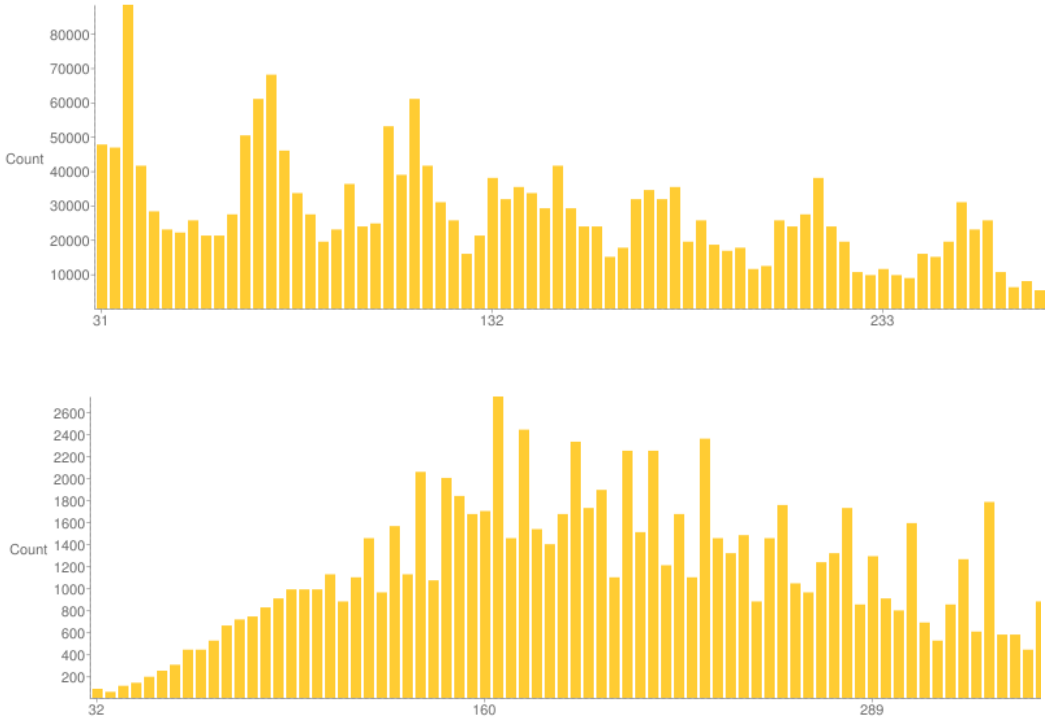


Figure 12. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for YNP bison aligned to Templeton.

Historic Bison

Historic sequences S6 and S9 had more sequence reads to compare to the bison reference sequence than wood buffalo and CCSP bison, , with an average of 203,802,966 reads, due to more allowed coverage during the sequencing process (Table 31). Historic sequences were first trimmed using FASTQ-MCF and then aligned to the bison reference sequence. Aligned sequencing statistics from the flagstaff option in Samtools for each historic sample can be found in Table 31.

Table 31. Samtools flagstat statistics of the 2 historic bison raw sequences mapped to Templeton (reads are in base pairs).

Statistic	S6	S9
in total (QC-passed reads + QC-failed reads)	281,220,542	126,385,390
duplicates	0	0
Mapped	280,139,485	119,801,801
(% mapped)	(99.62%)	(94.79%)
paired in sequencing	281,220,542	126,385,390
read1	139,634,378	62,828,857
read2	141,586,164	63,556,533
properly paired	255,124,934	104,192,008
with itself and mate mapped	280,022,221	119,778,264
singletons	117,264	23,537
with mate mapped to a different chr	25,524,820	16,317,742
with mate mapped to a different chr (mapQ>=5)	12,724,353	8,270,368

There were 11,857,832 SNVs and 246,878 INDELs, with 112,949 insertions and 134,501 deletions detected between historic sample S6 and the bison reference sequence (Table 32). There were 6,635,219 SNVs and 85,197 INDELS, with 35,791 insertions and 49,406 deletions identified between historic sample S9 and the bison reference

sequence (Table 32). In total there were 12,105,282 and 6,720,416 variants detected between historic sample S6 and S9 and the bison reference sequence, respectively.

Most of these variants were found to be heterozygous for one of the bison reference alleles and a variant allele. For historical sample S6 the percentage of reference variants found were approximately 75.7% and 52.5% for SNVs and INDELs, respectively. For historical sample S9 the percentage of reference alleles found for SNVs and INDELs were 76.8% and 68.3%, respectively. The homozygous variants

Historic samples were combined to do genomic variants comparison between S6 and S9 and to obtain statistics for the combined variants. The main base substitution was a G>A transition, with 3,797,976 detected (Table 33). The calculated transition to transversion ratio was 1.81 and 2.06 for S6 and S9, respectively (Table 34). The Ts/Tv ration for S6 was lower than the desired value of 2.0 like previous samples in this study, but quality of the variants in the analysis suggests that they were above the cutoff value of 30 and are most likely real. When combined with S9, the Ts/Tv ration did increase to 1.9. Variant counts and corresponding quality scores for both historical samples S6 and S9 can be found in figures 13 and 14, respectively.

Table 32. Summary statistics of historical bison samples variants detected from alignment to Templeton.

	S6		S9	
	SNVs	INDELs	SNVs	INDELs
Homozygous Variant alleles	2,874,357	116,763	1,535,659	26,707
Heterozygous (one Reference one variant)	8,977,673	129,539	5,093,918	57,988
Variant Count	11,857,832	246,878	6,632,398	84,947
Reference Alleles	8,977,673	129,539	5,093,918	57,988
Unique Alleles	11,124,271	235,245	5,898,837	73,314
Heterozygous Variant Alleles	5,802	576	2,821	252

Table 33. Base substitution counts for historical bison from alignment to Templeton.

	A	C	G	T
A	0	629,109	2,258,823	622,022
C	1,369,972	0	573,432	3,729,691
G	3,797,976	570,564	0	1,153,495
T	914,064	2,255,501	624,204	0

Table 34. Transition and transversion counts and ratio for each historic bison sample found against Templeton.

	S6	S9	Total
Transitions	9,508,678	5,504,782	15,013,460
Transversions	5,229,313	2,666,096	7,895,409
Ts/Tv ratio	1.81	2.06	1.9

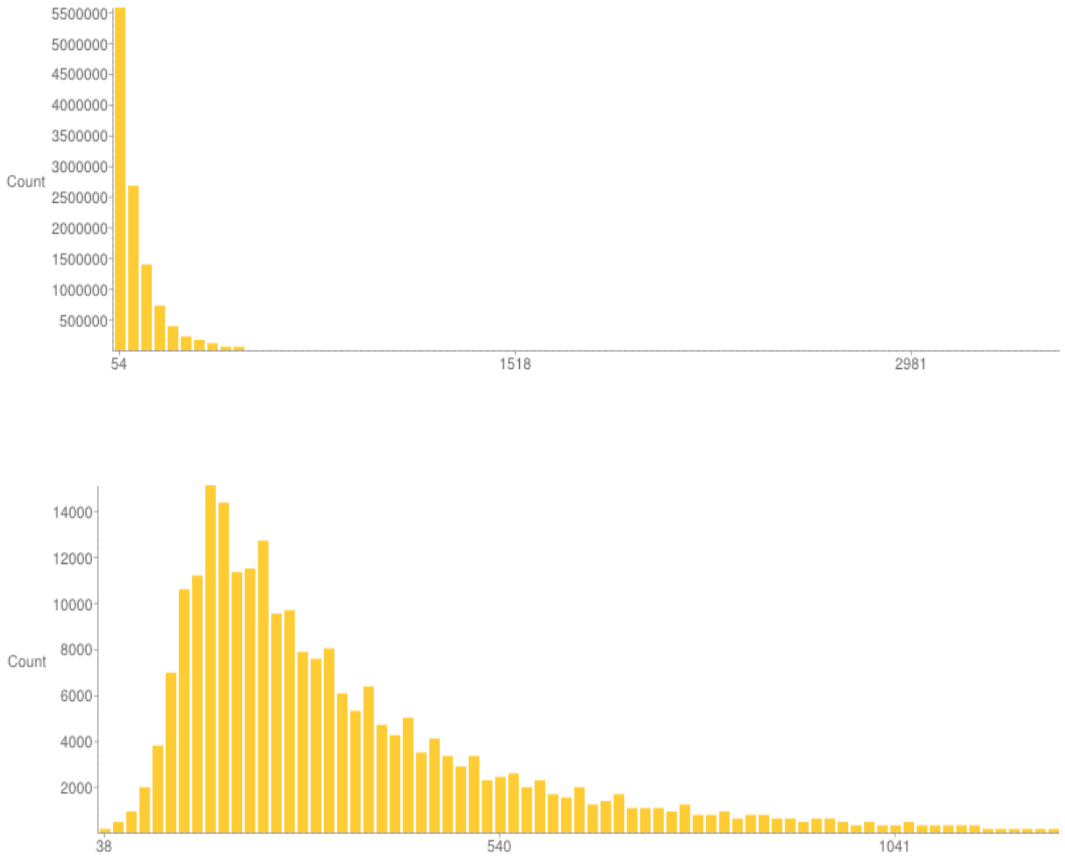


Figure 13. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for historic bison sample S6.

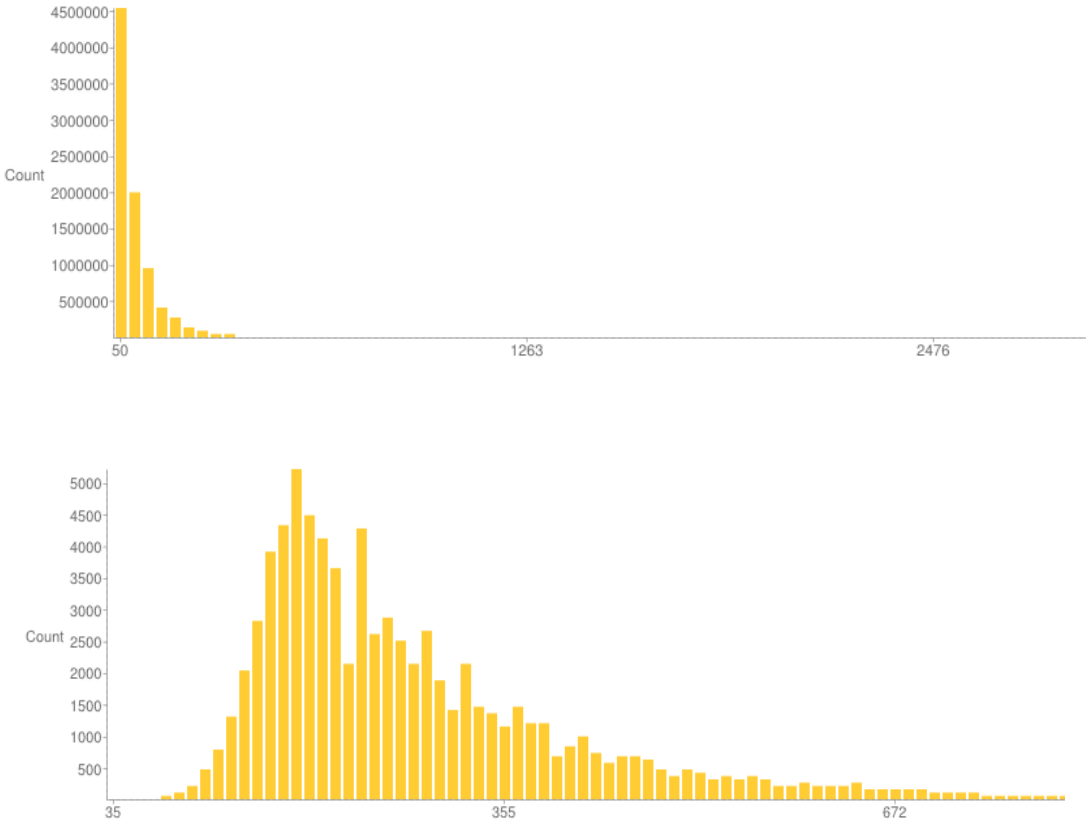


Figure 14. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for historic bison sample S9.

Comparison of All Bison Sample Variants to Templeton

For the analysis comparison of all bison variants found to Templeton, the population (CCSP, EIW, YNP, and Historical Samples) variant type tables were compared to identify those populations with the most variants detected (Table 35). As stated previously those samples with higher sequencing coverage (YNP and historical samples) were found to have more variants detected, both SNVs and INDELs, to Templeton than those populations with less sequencing coverage (CCSP and Wood).

After combining the common variants tables between populations we can better see which populations have the most informative SNPs and INDELs (Table 36). Since these populations were representing different sub-species of bison, they were not combined together to identify the total amount of informative variants that would be found for all 14 bison. Instead Table 36 offers informative variants found for each population and there corresponding number of individuals. Therefore, we can see that CCSP has 103,125 informative SNPs to Templeton and EIW only has 21,683 informative SNPs. The fewer informative SNPs for EIW could be due to the 2 samples having fewer reads for comparison to Templeton, which if we look at the number of SNPs in common for only 2 EIW samples there are 251,115 informative SNPs. The 4 YNP samples were found to have 741,721 informative SNPs detected when aligned to Templeton. The historic samples, S6 and S9 were combined to determine informative SNPs to Templeton and 733,561 SNPs were found to be informative for the historic samples. These informative SNPs between populations can be combined with unique alleles between populations to verify taxonomic status of these bison populations. The unique variants found for each population can be placed into a database for future validation of other animals from these populations.

Table 35. Summary of detected variants after alignment to Templeton by population.

Variant Type	CCSP	EIW	YNP	S6	S9
SNPs	3,877,737	2,192,618	9,171,543	11,857,832	6,635,219
INSERTIONs	14,769	6,408	208,771	112,949	35,791
DELETIONs	13,683	5,593	202,350	134,501	49,406
Total	3,906,189	2,204,619	9,582,664	12,105,282	6,720,416

Table 36. Number of common variants to Templeton found between all individuals by populations.

	CCSP		EIW		YNP		Historic Samples	
Shared Between	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs
4	103,125	1,394	21,683	496	741,721			
1	2,577,386	24,106	1,879,071	10,109	3,834,283	292,466	17,023,108	308,559
3	289,403	749	37,500	400	1,969,047	22,802		
2	902,866	2,084	251,115	926	2,612,899	93,107	733,561	11,633
Count	3,872,780	28,333	2,189,369	11,931	9,157,950	408,375	17,756,669	320,192

Table 37. Ratios of variants identified for all 4 populations to Templeton.

	CCSP		EIW		YNP		Historic Samples	
	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs
Homozygous Variant alleles	0.773	0.8308	0.7544	0.7689	0.4500	0.6996	0.237	0.3937
Heterozygous (one Reference one variant)	0.2267	0.164	0.2451	0.2225	0.5497	0.2999	0.7626	0.6037
Reference Alleles	0.2267	0.164	0.2451	0.2225	0.5497	0.2999	0.7626	0.6037
Unique Alleles	0.3718	0.6228	0.402	0.3434	0.2899	0.6148		
Heterozygous Variant Alleles	0.0003	0.0052	0.0005	0.0086	0.0003	0.0005	0.0005	0.0026

To compare the populations further, ratios of homozygous, heterozygous and reference alleles were calculated for each of the 4 population's genomic variants (Table 37). The historic and YNP samples were found to have higher percentages of reference alleles for SNPs than CCSP and EIW. Since the historic and YNP samples came from the same area, and represent northern plains bison it is expected that they would have more reference alleles than CCSP and EIW. CCSP was found to have the fewest reference alleles and the highest amount of homozygous variants, which could be due to being a closed population for so long.

To take into consideration on over-estimation of homozygous variants due to the lower coverage samples, the combined population VCF file was filtered based on QD and DP, following the VCF-tools filter options. In order to determine if filtering by coverage would change the number of total variants and type of variants (homozygous and heterozygous) the VCF with all of the samples was filtered at varying coverages, 5X, 10X, 20X for allele calls. The same trend (or ratio) of homozygous to heterozygous (reference) was found for all the samples and depending on the coverage used for filtering only a small amount of variants were lost. Using a less strict method of calling variants for further downstream analysis allowed for a broader amount of genomic variants to be analyzed, to ensure that detected variants were not removed from annotation. However, future validation still needs to be done for these variants reported in this research.

Variant Annotation to Templeton

In order to annotate the detected SNPs of the 14 samples above, the scaffolds in the VCF file needed to be replaced by a chromosome number. Using the SyMap produced pseudochromosomes from Templeton (see Chapter 1), the bison scaffolds used for the alignment were anchored to positions on respective chromosomes. This allowed for the scaffolds in the combined VCF files for each population, or in the case of the historical samples individually, to be replaced by chromosome based on the positions created in the SyMap anchor file and using a perl script. The changed VCF files were than annotated in SnpEff using UMD3.1.76 as a reference since no bison reference is available for the SnpEff software and the pseudochromosomes were generated from synteny blocks to the UMD31.76 reference.

When comparing the identified variants for each population from either using the VCF files that contained Templeton's scaffolds or Templeton's pseudo-chromosomes, a reduced amount of variants detected can be seen (Table 38). Due to this reduction of variants to analyze for each population, the variants detected were analyzed both with Templeton's scaffolds (previous analysis results) and pseudo-chromosomes.

An average approximately 1,905,704 variants were analyzed for all the populations. The variants for each population were annotated to produce annotated

Gene ID lists for each population. They were then compared to determine which annotated Gene IDs were in common for the populations and historic samples and which ones had unique cases (i.e. not found in one or more populations). There were a total of 3,420 Ensemble gene IDs that were annotated from SnpEff when all of the populations were compared for their annotated gene lists. Of these 3,420 annotated Gene IDs, 3,303 (96.6%; Appendix C) were found to be in common in all populations and historic samples. EIW were found to not have variants detected for 64 of the 3,420 (1.87%) annotated genes, while CCSP were found to not have variants detected for 14 of the 3,420 (0.41%) annotated genes. EIW and CCSP were both found to not have variants annotated for 22 of the 3,420 (0.64%) Gene IDs. The Gene IDs for these special cases and others can be found in Table 40.

The number of variants found for each population on each pseudo-chromosome can be found in Figure 15. Chromosomes 3, 5, and 7 were found to have the most variants for the populations. The majority of the identified annotated genes were found to be protein coding genes for each population (Table 39).

Table 38. Comparison of variants identified for each population when analyzed from alignment to Templeton's scaffolds (S) or pseudo-chromosomes (C).

	CCSP		EIW		YNP		S6		S9	
Variant Type	S	C	S	C	S	C	S	C	S	C
SNPs	3,877,737	1,047,865	2,192,618	582,671	9,171,543	2,505,736	11,857,832	3,316,529	6,635,219	1,863,012
INSERTIONS	14,769	3,590	6,408	1,318	208,771	57,185	112,949	31,022	35,791	9,567
DELETIONS	13,683	3,318	5,593	1,287	202,350	54,586	134,501	37,200	49,406	13,636
Total	3,906,189	1,054,773	2,204,619	585,276	9,582,664	2,617,507	12,105,282	3,384,751	6,720,416	1,886,215

Table 39. Biological functions of genes associated with annotated SNPs for all 14 bison based on populations, or individuals for S6 and S9.

	CCSP		EIW		YNP		S6		S9	
Biological Type	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs
Ribosomal RNA	91	15	91	5	91	78	91	68	87	45
Miscellaneous RNA	14		11		13	6	14	8	14	4
Protein coding	2,731	1,002	2,710	528	2,755	2,432	2,736	2,357	2,731	1,756
Small nucleolar RNA	128	23	121	11	128	89	119	90	126	46
Pseudogene	89	15	86	7	89	75	90	70	89	11
Processed pseudogene	21	4	20	1	23	17	22	12	23	37
Micro RNA	123	14	116	7	125	101	124	87	123	50
Small nuclear RNA	172	27	167	14	179	125	180	106	173	62
Total	3,369	1,100	3,322	573	3,403	2,923	3,376	2,798	3,366	2,011

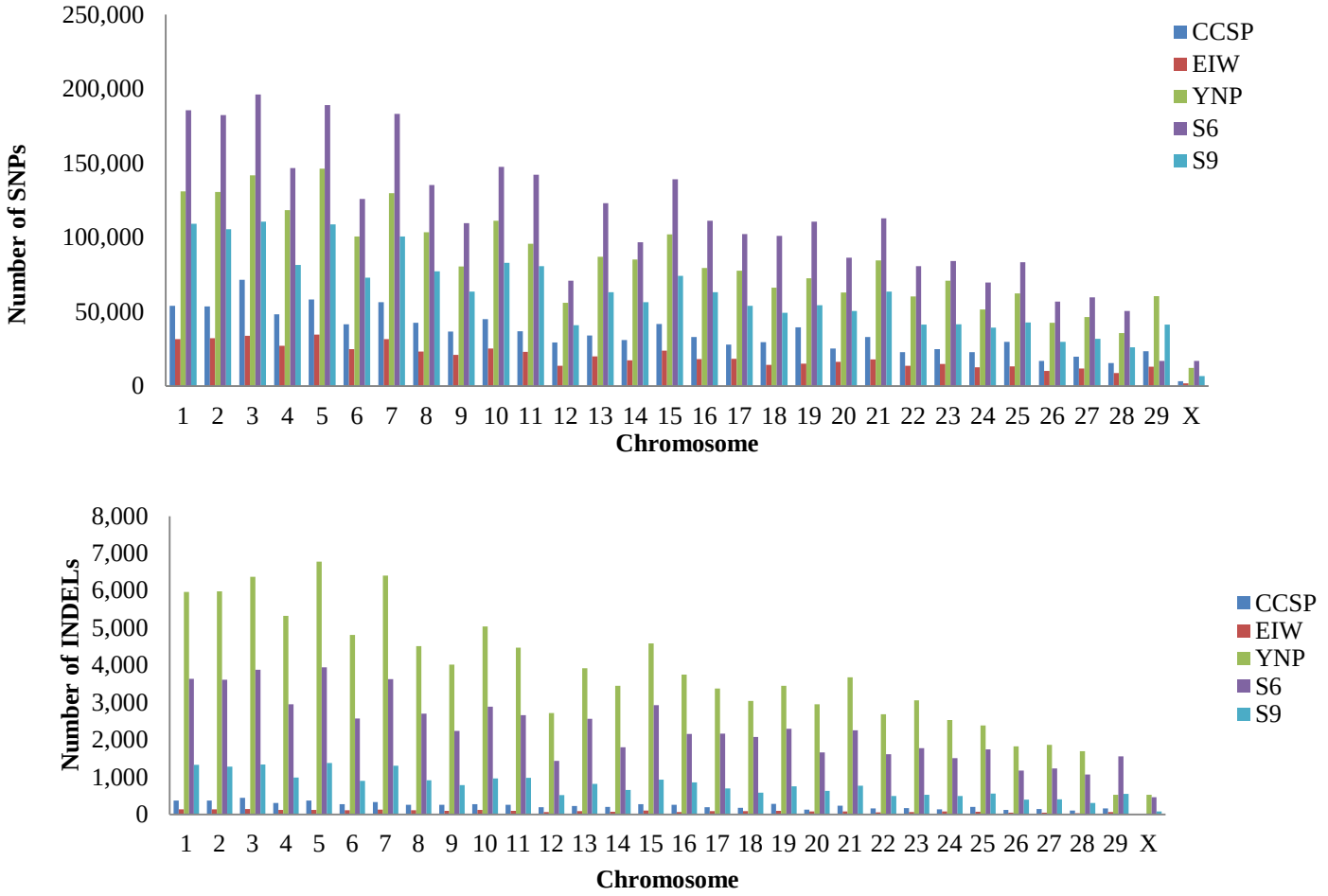


Figure 15. Number of SNPs (top) and INDELs (bottom) by chromosome for each population when using Templeton's pseudo-chromosomes.

Table 40. Unique genes after comparison of all bison genes found from annotation of variants identified to domestic cattle (see the notes section; abbreviations C=CCSP, Y=YNP, S6 and S9=historical samples and W=EIW).

Gene Symbol	Gene ID	Transcript ID	Biological Type	Notes
INTS9	ENSBTAG00000008845	ENSBTAT00000049079	protein_coding	S6S9 not YWC
ENSBTAG00000025379	ENSBTAG00000025379	ENSBTAT00000035638	pseudogene	S6S9 not YWC
SDR16C6	ENSBTAG00000040321	ENSBTAT00000052250	protein_coding	S6S9 not YWC
SNORA70	ENSBTAG00000044530	ENSBTAT00000061963	snoRNA	S6S9 not YWC
MRPL53	ENSBTAG00000016599	ENSBTAT00000022082	protein_coding	S6S9W not YC
EGR4	ENSBTAG00000024058	ENSBTAT00000033161	protein_coding	S6S9W not YC
UBL5	ENSBTAG00000040494	ENSBTAT00000022469	protein_coding	S6S9W not YC
RAB25	ENSBTAG00000018914	ENSBTAT00000025170	protein_coding	S6S9WC not Y
ATP5J	ENSBTAG00000000605	ENSBTAT00000032427	protein_coding	S6S9Y not WC
TEX261	ENSBTAG00000002105	ENSBTAT00000047005	protein_coding	S6S9Y not WC
ENSBTAG00000002655	ENSBTAG00000002655	ENSBTAT00000003439	protein_coding	S6S9Y not WC
PODNL1	ENSBTAG00000006073	ENSBTAT00000007981	protein_coding	S6S9Y not WC
SWSAP1	ENSBTAG00000008196	ENSBTAT00000010778	protein_coding	S6S9Y not WC
ANKRD53	ENSBTAG00000010623	ENSBTAT00000014047	protein_coding	S6S9Y not WC
PEX11G	ENSBTAG00000018894	ENSBTAT00000025148	protein_coding	S6S9Y not WC
NLE1	ENSBTAG00000019094	ENSBTAT00000025423	protein_coding	S6S9Y not WC
FAM49B	ENSBTAG00000020801	ENSBTAT00000040368	protein_coding	S6S9Y not WC
CASP13	ENSBTAG00000020884	ENSBTAT00000027820	protein_coding	S6S9Y not WC
POLR2K	ENSBTAG00000022539	ENSBTAT00000030510	protein_coding	S6S9Y not WC
ENSBTAG00000035021	ENSBTAG00000035021	ENSBTAT00000049500	processed_pseudogene	S6S9Y not WC
ENSBTAG00000038900	ENSBTAG00000038900	ENSBTAT00000054428	protein_coding	S6S9Y not WC
MARVELD2	ENSBTAG00000040001	ENSBTAT00000053818	protein_coding	S6S9Y not WC
U6	ENSBTAG00000042261	ENSBTAT00000059253	snRNA	S6S9Y not WC
U6	ENSBTAG00000042480	ENSBTAT00000059472	snRNA	S6S9Y not WC

Table 40. Continued

Gene Symbol	Gene ID	Transcript ID	Biological Type	Notes
U6	ENSBTAG00000043422	ENSBTAT00000060414	snRNA	S6S9Y not WC
U6	ENSBTAG00000044931	ENSBTAT00000062364	snRNA	S6S9Y not WC
U1	ENSBTAG00000046205	ENSBTAT00000064797	snRNA	S6S9Y not WC
ENSBTAG00000046655	ENSBTAG00000046655	ENSBTAT00000064108	protein_coding	S6S9Y not WC
ENSBTAG00000046904	ENSBTAG00000046904	ENSBTAT00000063790	processed_pseudogene	S6S9Y not WC
TSC22D1	ENSBTAG00000047739	ENSBTAT00000025431	protein_coding	S6S9Y not WC
FOXF1	ENSBTAG00000000009	ENSBTAT00000000009	protein_coding	S6S9YC not W
GTDC2	ENSBTAG00000000459	ENSBTAT00000000583	protein_coding	S6S9YC not W
PRCC	ENSBTAG00000000608	ENSBTAT00000000800	protein_coding	S6S9YC not W
CCDC130	ENSBTAG00000002516	ENSBTAT00000003269	protein_coding	S6S9YC not W
SLC25A23	ENSBTAG00000003491	ENSBTAT00000004536	protein_coding	S6S9YC not W
ENSBTAG00000004219	ENSBTAG00000004219	ENSBTAT00000005528	protein_coding	S6S9YC not W
CBLN2	ENSBTAG00000005985	ENSBTAT00000007856	protein_coding	S6S9YC not W
ENSBTAG00000007443	ENSBTAG00000007443	ENSBTAT00000009786	pseudogene	S6S9YC not W
CCDC151	ENSBTAG00000008201	ENSBTAT00000010785	protein_coding	S6S9YC not W
ENSBTAG00000008244	ENSBTAG00000008244	ENSBTAT00000010845	protein_coding	S6S9YC not W
TOM40B	ENSBTAG00000009213	ENSBTAT00000012142	protein_coding	S6S9YC not W
DHRS11	ENSBTAG00000010297	ENSBTAT00000013600	protein_coding	S6S9YC not W
CATB	ENSBTAG00000012442	ENSBTAT00000036795	protein_coding	S6S9YC not W
CRP	ENSBTAG00000013907	ENSBTAT00000018469	protein_coding	S6S9YC not W
ENSBTAG00000014365	ENSBTAG00000014365	ENSBTAT00000057152	protein_coding	S6S9YC not W
CCR8	ENSBTAG00000015483	ENSBTAT00000020577	protein_coding	S6S9YC not W
ENSBTAG00000015493	ENSBTAG00000015493	ENSBTAT00000020594	protein_coding	S6S9YC not W
C22H3ORF10	ENSBTAG00000016098	ENSBTAT00000021429	protein_coding	S6S9YC not W

Table 40. Continued

Gene Symbol	Gene ID	Transcript ID	Biological Type	Notes
KCNG4	ENSBTAG00000016695	ENSBTAT00000022193	protein_coding	S6S9YC not W
MRPL39	ENSBTAG00000019542	ENSBTAT00000026038	protein_coding	S6S9YC not W
ENSBTAG00000019618	ENSBTAG00000019618	ENSBTAT00000054917	protein_coding	S6S9YC not W
C29H11orf73	ENSBTAG00000019995	ENSBTAT00000026632	protein_coding	S6S9YC not W
ENSBTAG00000020765	ENSBTAG00000020765	ENSBTAT00000054562	protein_coding	S6S9YC not W
U1	ENSBTAG00000028119	ENSBTAT00000040501	snRNA	S6S9YC not W
bta-mir-138-1	ENSBTAG00000029814	ENSBTAT00000042193	miRNA	S6S9YC not W
bta-mir-365-1	ENSBTAG00000029848	ENSBTAT00000042227	miRNA	S6S9YC not W
PRR35	ENSBTAG00000033739	ENSBTAT00000056123	protein_coding	S6S9YC not W
ZNF35	ENSBTAG00000034005	ENSBTAT00000011532	protein_coding	S6S9YC not W
GNG11	ENSBTAG00000034449	ENSBTAT00000048797	protein_coding	S6S9YC not W
bta-mir-193b	ENSBTAG00000036371	ENSBTAT00000050871	miRNA	S6S9YC not W
ENSBTAG00000038094	ENSBTAG00000038094	ENSBTAT00000052085	miRNA	S6S9YC not W
ENSBTAG00000039104	ENSBTAG00000039104	ENSBTAT00000055189	miRNA	S6S9YC not W
ENSBTAG00000039858	ENSBTAG00000039858	ENSBTAT00000054277	miRNA	S6S9YC not W
FOXL1	ENSBTAG00000040605	ENSBTAT00000053216	protein_coding	S6S9YC not W
SNORD115	ENSBTAG00000042397	ENSBTAT00000059389	snoRNA	S6S9YC not W
SNORA70	ENSBTAG00000042712	ENSBTAT00000059704	snoRNA	S6S9YC not W
SNORD115	ENSBTAG00000042769	ENSBTAT00000059761	snoRNA	S6S9YC not W
U6	ENSBTAG00000043253	ENSBTAT00000060245	snRNA	S6S9YC not W
U6	ENSBTAG00000043254	ENSBTAT00000060246	snRNA	S6S9YC not W
U6	ENSBTAG00000043427	ENSBTAT00000060419	snRNA	S6S9YC not W
SNORD115	ENSBTAG00000043599	ENSBTAT00000060591	snoRNA	S6S9YC not W
SNORA35	ENSBTAG00000043696	ENSBTAT00000060688	snoRNA	S6S9YC not W

Table 40. Continued

Gene Symbol	Gene ID	Transcript ID	Biological Type	Notes
MAF	ENSBTAG00000044192	ENSBTAT00000061511	protein_coding	S6S9YC not W
U6	ENSBTAG00000044237	ENSBTAT00000061670	snRNA	S6S9YC not W
7SK	ENSBTAG00000044283	ENSBTAT00000061716	misc_RNA	S6S9YC not W
7SK	ENSBTAG00000044586	ENSBTAT00000062019	misc_RNA	S6S9YC not W
U6	ENSBTAG00000044988	ENSBTAT00000062421	snRNA	S6S9YC not W
ENSBTAG00000045072	ENSBTAG00000045072	ENSBTAT00000062505	miRNA	S6S9YC not W
ENSBTAG00000045087	ENSBTAG00000045087	ENSBTAT00000062520	miRNA	S6S9YC not W
SNORA70	ENSBTAG00000045173	ENSBTAT00000062606	snoRNA	S6S9YC not W
SCARNA15	ENSBTAG00000045209	ENSBTAT00000062642	snoRNA	S6S9YC not W
ENSBTAG00000045517	ENSBTAG00000045517	ENSBTAT00000064447	pseudogene	S6S9YC not W
U6	ENSBTAG00000045725	ENSBTAT00000062785	snRNA	S6S9YC not W
ENSBTAG00000046074	ENSBTAG00000046074	ENSBTAT00000064906	protein_coding	S6S9YC not W
ZNF705A	ENSBTAG00000046204	ENSBTAT00000065604	protein_coding	S6S9YC not W
ENSBTAG00000046233	ENSBTAG00000046233	ENSBTAT00000064725	protein_coding	S6S9YC not W
ENSBTAG00000046335	ENSBTAG00000046335	ENSBTAT00000063057	processed_pseudogene	S6S9YC not W
ENSBTAG00000046689	ENSBTAG00000046689	ENSBTAT00000065792	miRNA	S6S9YC not W
ENSBTAG00000046789	ENSBTAG00000046789	ENSBTAT00000063503	protein_coding	S6S9YC not W
OR10J1	ENSBTAG00000047063	ENSBTAT00000064894	protein_coding	S6S9YC not W
Metazoa_SRP	ENSBTAG00000047359	ENSBTAT00000065479	misc_RNA	S6S9YC not W
ENSBTAG00000047369	ENSBTAG00000047369	ENSBTAT00000065787	pseudogene	S6S9YC not W
ENSBTAG00000047555	ENSBTAG00000047555	ENSBTAT00000065396	protein_coding	S6S9YC not W
U6	ENSBTAG00000047929	ENSBTAT00000063210	snRNA	S6S9YC not W
ENSBTAG00000000560	ENSBTAG00000000560	ENSBTAT00000000730	protein_coding	S6S9YW not C
RB1CC1	ENSBTAG00000000878	ENSBTAT00000001169	protein_coding	S6S9YW not C

Table 40. Continued

Gene Symbol	Gene ID	Transcript ID	Biological Type	Notes
NSA2	ENSBTAG00000003066	ENSBTAT00000003989	protein_coding	S6S9YW not C
LSM6	ENSBTAG00000014060	ENSBTAT00000018681	protein_coding	S6S9YW not C
ARNT	ENSBTAG00000021037	ENSBTAT00000028023	protein_coding	S6S9YW not C
PGPEP1L	ENSBTAG00000021531	ENSBTAT00000028693	protein_coding	S6S9YW not C
SNN	ENSBTAG00000026376	ENSBTAT00000037468	protein_coding	S6S9YW not C
ENSBTAG00000039983	ENSBTAG00000039983	ENSBTAT00000055185	protein_coding	S6S9YW not C
U6atac	ENSBTAG00000042146	ENSBTAT00000059138	snRNA	S6S9YW not C
U6	ENSBTAG00000043505	ENSBTAT00000060497	snRNA	S6S9YW not C
ENSBTAG00000045057	ENSBTAG00000045057	ENSBTAT00000062490	miRNA	S6S9YW not C
bta-mir-2456	ENSBTAG00000045270	ENSBTAT00000062703	miRNA	S6S9YW not C
ENSBTAG00000046727	ENSBTAG00000046727	ENSBTAT00000062985	protein_coding	S6S9YW not C
ENSBTAG00000047796	ENSBTAG00000047796	ENSBTAT00000053325	protein_coding	S6S9YW not C
U5	ENSBTAG00000043085	ENSBTAT00000060077	snRNA	S6So not S9YWC
DDX58	ENSBTAG00000003366	ENSBTAT00000061377	protein_coding	S6So not S9YWC
PI4KB	ENSBTAG00000007320	ENSBTAT00000009627	protein_coding	S6So not S9YWC
ZNF511	ENSBTAG00000012416	ENSBTAT00000016475	protein_coding	S6So not S9YWC
U1	ENSBTAG00000029683	ENSBTAT00000042062	snRNA	S6So not S9YWC
ENSBTAG00000034845	ENSBTAG00000034845	ENSBTAT00000049313	protein_coding	S6So not S9YWC
U6	ENSBTAG00000042563	ENSBTAT00000059555	snRNA	S6So not S9YWC
U6	ENSBTAG00000044554	ENSBTAT00000061987	snRNA	S6So not S9YWC
ENSBTAG00000046202	ENSBTAG00000046202	ENSBTAT00000050512	protein_coding	YWCS not S6S9

The 3,420 gene IDs were ran in DAVID to do a Gene Ontology (GO) analysis and to determine enriched biological, and were found to match 2,691 DAVID IDs in the *Bos taurus* database. Choosing only the gene ontology for biological pathway FAT option and using the Functional Annotation Chart, 89 chart records were produced for biological pathways and consisted of 1,049 genes from the imported list. Only 50 enriched gene ontology categories for biological pathways were produced with a Fisher Exact P-value ≤ 0.05 but had higher false discovery values (Appendix D). These 50 GO terms were mainly associated with regulatory functions in domestic cattle.

Discussion

With using Illumina paired-end resequenced bison raw reads and aligning them to the bison reference genome we were able to identify new genomic variants among different bison populations. This multi-way comparison of bison genome sequences was used to determine genomic differences between each population and the reference population, as well as comparing these differences across populations.

EIW buffalo were found to mainly have homozygous variants when aligned to Templeton, and also the second highest homozygous variant ratio when compared to the other 3 bison populations. Even though EIW buffalo are classified as a different sub-species than Templeton, we did see some reference alleles in the variants called between the 4 samples. Using the variants that were found between Templeton and EIW we can determine the percent of the wood buffalo genome that is shared with plains bison and those regions that are found to be wood buffalo specific in future studies to describe

evolutionary differences between wood buffalo and plains bison to help determine taxonomic status of bison sub-species.

For the CCSP bison we detected more homozygous variants when compared to the other bison populations, which could be resultant of the closed herd management practices the herd has been under for years. The CCSP bison were found to possess their own unique alleles when compared to the other bison populations. These unique alleles could define the genomic remnants of the southern plains bison genetics in CCSP. Further analysis will be needed to verify these unique allele's in future re-sequenced CCSP bison to evaluate the taxonomic status of the CCSP bison and their ancestry to the last animals of the southern plains bison.

The YNP variants detected were primarily heterozygous variants that possessed a reference allele. With the YNP samples coming from the same herd as Templeton it is expected to see some reference alleles and fewer variants for these samples. The fact that there was on average 44% of the variants being homozygous between the 4 samples and Templeton suggests that even samples from the same population can have a significant amount of genomic variants between them. The population dynamics of YNP could influence these variations by the larger population numbers that exist when compared to other bison herds.

Lastly, the historic bison samples provided the most heterozygous variants to Templeton and the fewest homozygous variants detected. Those variants that are homozygous to Templeton can be used to identify parts of the genome that have been possibly been lost over time to describe evolutionary differences between historic and

modern bison in future studies. The historic samples were not compared to the other 3 bison populations for unique variants, but rather to each other. This comparison allows us to verify that these variants detected were not sequencing error but can be found in both historic samples. This identified 733,561 and 11,633 informative SNPs and INDELs, respectively. These informative variants can be placed into a database for future comparison of modern bison to evaluate if any bison possess these unique variants.

Those alleles that were identified as unique alleles for certain individuals need to be validated before they can be designated as an informative variant for certain populations. This can be done by creating a dataset of these variants and then with future bison sequences analyzed and comparing the variants detected to this dataset. With future re-sequenced bison samples we can validate these identified variants to be able to provide a genomic data set with known variants between different bison populations for bison conservation management.

Variant Identification to Domestic Cattle

EIW

The raw Illumina paired-end sequences of 4 wood buffalo were individually aligned to the domestic cattle UMD3.1 (Ensembl GCA_000003055.3) sequence using the mem alignment option of BWA (same as above to the bison reference). The SAMtools options, view and flagstat, were used to obtain statistics of the 4 wood buffalo Illumina paired-end reads mapped to the domestic cattle reference sequence. As above, the same two EIW samples (sample IDs 95-1573 and 151-1607) were found to have fewer reads to be aligned to the domestic cattle reference sequence. Comparisons and statistics for each EIW sample can be found in Table 41.

Table 41. Samtools flagstat statistics of the 4 EIW bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 (reads are in base pairs).

Statistic	26-1525	95-1573	151-1607	233-1676
in total (QC-passed reads + QC-failed reads)	58,711,530	12,968,260	17,859,638	75,729,836
duplicates	0	0	0	0
Mapped	37,336,337	8,675,298	12,826,096	39,001,290
(% mapped)	(63.59%)	(66.90%)	(71.82%)	(51.50%)
paired in sequencing	58,711,530	12,968,260	17,859,638	75,729,836
read1	29,355,765	6,484,130	8,929,819	37,864,918
read2	29,355,765	6,484,130	8,929,819	37,864,918
properly paired	30,827,523	6,905,076	10,599,882	32,206,127
with itself and mate mapped	33,441,388	7,250,197	11,140,290	35,614,310
singletons	3,894,949	1,425,101	1,685,806	3,386,980
with mate mapped to a different chr	1,845,884	293,737	469,254	1,795,594
with mate mapped to a different chr (mapQ>=5)	810,414	128,044	205,837	761,582

SNVs and INDELs were called separately using GATK after the alignment of the raw Illumina reads to the domestic cattle reference. The analysis for variants will remain separate for SNVs and INDELs to give statistics of the different genomic variants identified. Table 42 offers a summary for each individual sample and their SNVs detected, where the most SNVs detected were found for wood buffalo 26-1525 with 5,138,270 SNVs. Most of the SNVs for each individual were found to be homozygous for the variant allele, with the fewest SNVs that were heterozygous for 2 different variant alleles. Reference alleles resulted from those SNVs that were found to be heterozygous for a reference and variant allele.

SAMtools vcf-stats and SnpEff were used to analyze each individual VCF files and combined VCF file for the wood buffalo population for statistics of SNVs and INDELs detected. A total of 9,590,819 total SNPs found between all 4 wood buffalo samples and the domestic cattle reference sequence (Table 43). Of these 9,590,819, only 26,494 (0.2%) SNPs were found to be in common between all 4 wood buffalo samples and 7,753,682 (80.8%) SNPs were at least found in common between two of the wood buffalo samples (Table 44). Overall there was one SNP detected within the wood population every 277 bases with 10,773,810 number of effects. The most common variant between wood buffalo and domestic cattle was G>A substitution at 1,680,391, with the least common substitution is an A>T with only 323,601 detected (Table 45). The transitions to transversions ratio was calculated for all EIW samples at 2.25, from 6,640,573 transitions and 2,953,891 transversions detected (Table 46). Table 46 also has

each individual Ts/Tv ratio for each EIW sample, and variants were detected with few errors.

Unique alleles are those variants that were detected in only one individual and not the other three when statistics were ran on the combined VCF population. There was a considerable amount of unique alleles that were detected for each individual wood buffalo sample. This could be due to sequencing errors or lower quality variants being called, which is taken into account for the annotation. The same was found for the INDELS. The comparison of individual INDEL files for each wood buffalo sample shows the same trend of unique alleles (Table 42). Chromosomal variant counts for both SNPs and INDELs for wood buffalo aligned to domestic cattle are depicted in Figure 16, with most of the variants found on Chromosome 1 for both type of variants.

Table 42. Individual variant summary statistics of 4 EIW buffalo found from comparing sequences to domestic cattle UMD3.1.

	26-1525		95-1573		151-1607		233-1676	
	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs
Homozygous Variant alleles	4,860,695	17,815	541,779	579	1,149,279	894	4,538,574	27,816
Heterozygous (one Reference one variant)	277,076	1,352	20,155	173	41,336	266	239,590	1,780
Variant Count	5,138,270	19,186	562,043	754	1,190,782	1,166	4,778,608	29,617
Reference Alleles	277,076	1,352	20,155	173	41,336	266	239,590	1,780
Unique Alleles	3,590,578	16,852	252,052	261	601,396	514	3,309,656	27,295
Heterozygous Variant Alleles	499	19	109	2	167	6	444	21

Table 43. Summary statistics for SNPs and INDELs found in 4 EIW buffalo compared to domestic cattle UMD3.1.

	SNPs	INDELs
Warnings	959,569	5,899
Errors	63,885	757
Number of lines (input file)	9,590,819	47,457
Number of variants	9,594,464	47,526
Number of multi-allelic VCF entries	3,622	64
Number of effects	10,773,810	55,774
Genome total length	2,670,424,254	2,670,422,689
Genome effective length	2,660,908,360	2,660,906,795
Variant rate	1 variant every 277 bases	1 variant every 55,988 bases

Figure 17 shows the count of variants with corresponding quality scores of those SNVs and INDELs annotated with SnpEff after filtering, respectively. Therefore, we can see what quality scores are left of the variants annotated to see if quality scores are still qualified for further analysis.

There were 47,457 INDELs discovered between wood buffalo and domestic cattle, with 22,994 (48.5%) insertions and 24,532 (51.5%) deletions. All INDELs were used for annotation and 1 INDEL was detected every 55,988 bases, with 55,774 effects detected (Table 43). Only 177 (0.3%) of the INDELs that were found in common of all 4 wood buffalo samples, while 44,922 (94.6%) were found in at least two of the wood buffalo samples

Table 44. Number of common variants found between the 4 EIW samples.

Shared between	SNPs	INDELs
4	26,494	177
1	7,753,682	44,922
3	188,759	377
2	1,621,884	1,981
Total	9,590,819	47,457

Table 45. Base changes (SNPs) between EIW buffalo and domestic cattle UMD3.1.

	A	C	G	T
A	0	384,709	1,643,509	323,601
C	406,521	0	360,459	1,674,468
G	1,680,391	362,188	0	408,462
T	323,653	1,642,205	384,298	0

Table 46. Transition and transversion counts and ratio for each EIW sample and population totals found against domestic cattle reference UMD3.1.

Sample	48-5795	50-5792	61-5793	68-5784	Total
Transitions	8,535,821	9,888,021	7,555,401	7,710,099	6,640,573
Transversions	3,751,512	4,390,577	3,345,246	3,447,405	2,953,891
Ts/Tv	2.28	2.25	2.26	2.24	2.25

Summary of effects for SNPs and INDELs with SnpEff provides the number of effects by type and region (Table 47). For both SNPs and INDELs the majority of the variants were found in intergenic regions at 63.71% and 57.61%, respectively. The types of effects and percentages for each can be found for the SNPs and INDELs in

Appendices E and F. The biological functions of genes associated with the annotated SNPs and INDELs for wood buffalo where mainly protein coding genes for both variant types (Table 48). There were a total of 24,501 genes annotated from the wood buffalo SNPs and INDELS gene file. Of those only 14,996 (61.2%) were found in only the annotated SNPs file, compared to only 9,505 (38.8%) genes that were found from both the annotated SNV and INDEL gene files.

Table 47. Genomic regions associated with annotated SNPs and INDELs for combined EIW samples.

	SNPs		INDELs	
Region Type (alphabetical order)	Count	Percent	Count	Percent
DOWNSTREAM	413,868	3.84%	3,035	5.44%
EXON	65,706	0.61%	86	0.15%
INTERGENIC	6,864,044	63.71%	32,132	57.61%
INTRON	2,907,337	26.99%	15,978	28.65%
NONE	63,885	0.59%	757	1.36%
SPLICE_SITE_ACCEPTOR	167	0.00%	10	0.02%
SPLICE_SITE_DONOR	169	0.00%	5	0.01%
SPLICE_SITE_REGION	6,639	0.06%	45	0.08%
TRANSCRIPT	15	0%		
UPSTREAM	425,776	3.95%	3,407	6.11%
UTR_3_PRIME	21,229	0.20%	270	0.48%
UTR_5_PRIME	4,975	0.05%	49	0.09%

Table 48. Biological functions of genes associated with annotated SNPs and INDELs in EIW buffalo.

Biological Type	SNPs	INDELs
Protein coding	19,924	8,701
Ribosomal RNA	399	103
Miscellaneous RNA	175	15
Small nucleolar RNA	846	136
Pseudogene	621	127
Processed pseudogene	168	42
Micro RNA	1,151	197
Small nuclear RNA	1,217	184
Total	24,501	9,505

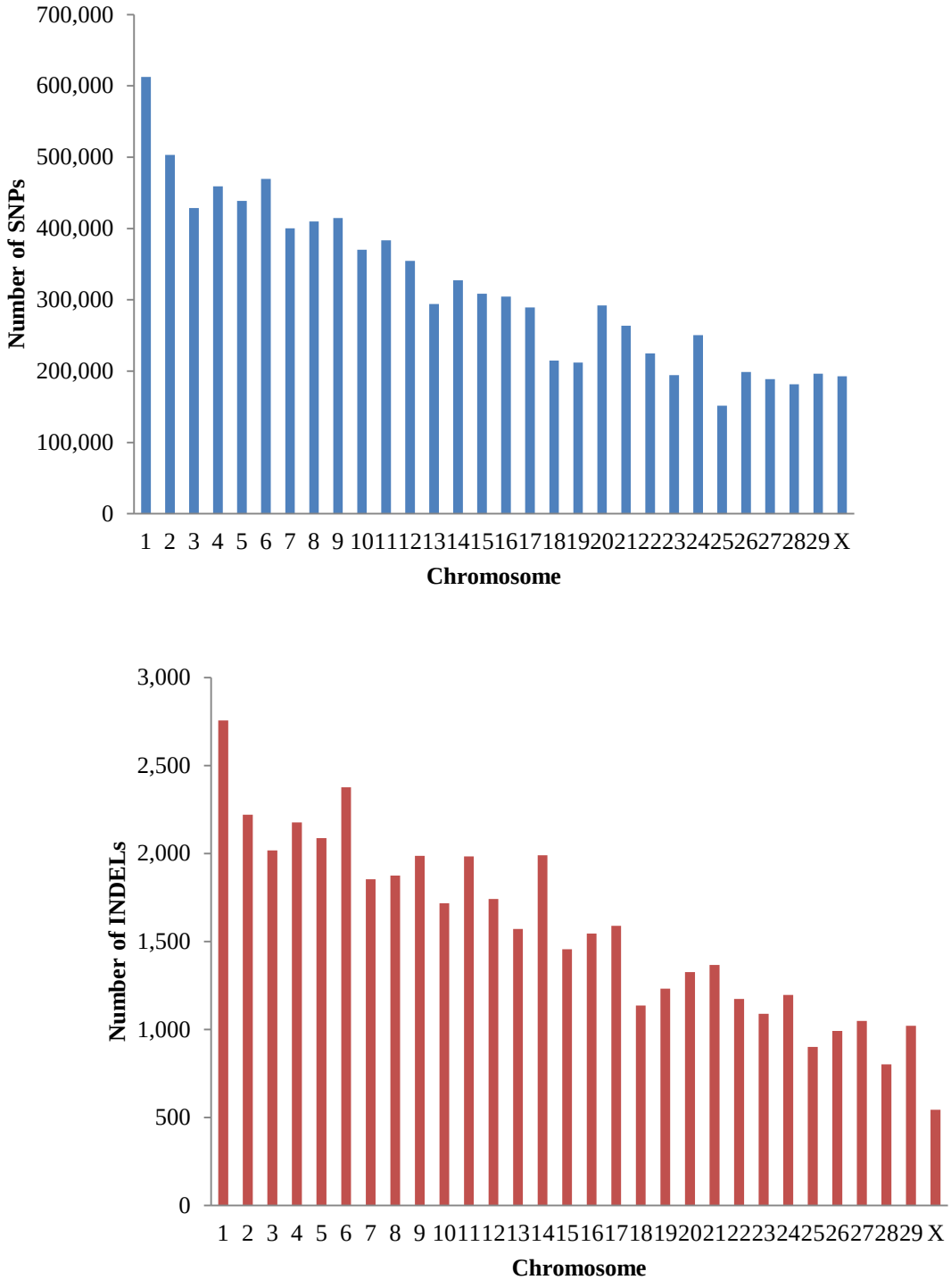


Figure 16. Number of SNPs (blue) and INDELs (red) by chromosome for EIW samples mapped to domestic cattle UMD3.1.

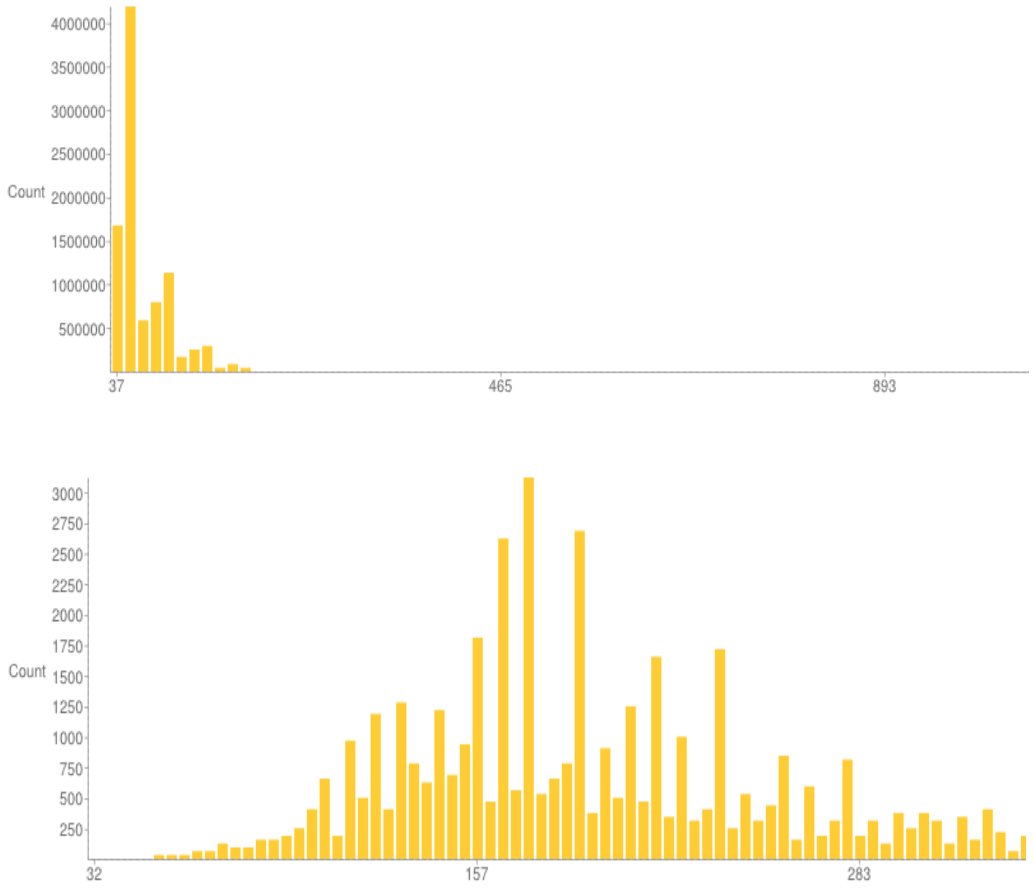


Figure 17. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for EIW buffalo.

CCSP

Raw Illumina paired-end sequences of 4 CCSP were individually aligned to the domestic cattle UMD3.1 sequence using the mem alignment option of BWA (same as above to the bison reference). The analysis for variants will remain separate for SNVs and INDELs and then grouped together to consider the variants within the population as a whole, the same as the results for wood buffalo.

Using the SAMtools option flagstat read statistics were obtaines for the 4 CCSP Illumina paired-end reads mapped to the domestic cattle reference sequence (Table 49). There was an average of 69,206,349 raw Illumina parried-end reads for the CCSP bison to be compared to the domestic cattle reference sequence.

Table 49. Samtools flagstat statistics of the 4 CCSP bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 (reads are in base pairs).

Statistic	48-5795	50-5792	61-5793	68-5784
in total (QC-passed reads + QC-failed reads)	74,724,288	71,988,836	66,683,988	63,428,284
duplicates	0	0	0	0
Mapped	47,504,785	47,979,920	40,616,520	40,226,147
(% mapped)	(63.57%)	(66.65%)	(60.91%)	(63.42%)
paired in sequencing	74,724,288	71,988,836	66,683,988	63,428,284
read1	37,362,144	35,994,418	33,341,994	31,714,142
read2	37,362,144	35,994,418	33,341,994	31,714,142
properly paired	37,486,507	40,869,798	33,451,407	33,627,053
with itself and mate mapped	40,106,536	44,032,334	36,194,157	36,225,709
singletons	7,398,249	3,947,586	4,422,363	4,000,438
with mate mapped to a different chr	1,852,134	2,148,064	1,732,905	1,659,771
with mate mapped to a different chr (mapQ>=5)	756,025	907,997	718,414	714,277

All of the samples were found to have similar amounts of SNPs and INDELs called, meaning that they all had similar sequence data to be analyzed for the analysis, unlike the EIW data. Table 50 offers a summary for each individual sample and their SNVs detected, where the most SNVs detected were found for CCSP 50-5792 with 6,951,488 SNVs. Most of the SNVs for each individual were found to be homozygous for the variant allele, with the fewest SNVs that were heterozygous for 2 different variant alleles. Reference alleles resulted from those SNVs that were found to be heterozygous for a reference and variant allele.

SAMtools vcf-stats and SnpEff were used to analyze each individual VCF files and combined VCF file for the CCSP population for statistics of SNVs and INDELs detected. A total of 15,625,724 total SNPs were annotated from the CCSP bison and the domestic cattle reference sequence (Table 51). Of these 15,625,724 SNPs, 407,095 (2.6%) SNPs were found to be in common between all four CCSP samples and 4,651,727 (29.8%) SNPs were at least found in common between two of the CCSP samples (Table 52). Overall there was one SNP detected within the CCSP population every 170 bases with 17,459,382 number of effects (Table 51). The most common variant between CCSP and domestic cattle was G>A substitution at 2,743,063, with the least common substitution is a T>A with only 525,950 detected (Table 52). The transitions to transversions ratio was calculated at 2.25, from 10,825,038 transitions and 4,800,686 transversions detected for the CCSP population (Table 54).

Table 50. Individual variant summary statistics of 4 CCSP found from comparing sequences to domestic cattle UMD3.1.

	68-5784		48-5795		61-5793		50-5792	
	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs	SNPs	INDELs
Homozygous Variant alleles	5,457,063	24,358	5,987,317	39,840	5,315,307	22,346	6,951,488	38,588
Heterozygous (one Reference one variant)	242,400	1,371	311,381	2,289	268,853	1,498	374,280	2,096
Variant Count	5,699,952	25,745	6,299,357	42,153	5,584,750	23,864	7,326,439	40,705
Reference Alleles	242,400	1,371	311,381	2,289	268,853	1,498	374,280	2,096
Unique Alleles	1,911,454	20,278	2,294,673	34,963	1,849,616	17,909	2,793,563	33,404
Heterozygous Variant Alleles	489	16	659	24	590	20	671	21

Table 51. Summary statistics for SNPs and INDELs found in 4 CCSP bison compared to domestic cattle.

	SNPs	INDELs
Warnings	1,511,826	12,581
Errors	85,566	1,053
Number of lines (input file)	15,617,914	117,201
Number of variants	15,625,724	117,308
Number of multi-allelic VCF entries	7,771	103
Number of effects	17,459,382	133,931
Genome total length	2,670,424,426	2,670,422,800
Genome effective length	2,660,908,532	2,660,906,906
Variant rate	1 variant every 170 bases	1 variant every 22,683 bases

There were 117,201 INDELs discovered between CCSP bison and domestic cattle, with 55,773 (47.6%) insertions and 61,535 (52.4%) deletions. All INDELs were used for annotation and 1 INDEL was detected every 22,683 bases, with 133,931 effects detected (Table 54). Only 1,426 (1.2%) of the INDELs that were found in common of all 4 CCSP bison samples, while 106,554 (90.9%) were found in at least two of the CCSP bison samples. Chromosomal variant counts for both SNPs and INDELs for CCSP

aligned to domestic cattle are depicted in Figure 18, with most of the variants found on Chromosome 1 for both type of variants. Figure 19 shows the count of quality scores for variants of those SNVs and INDELs annotated with SnpEff after filtering, respectively.

Table 52. Number of common variants found between the 4 CCSP samples.

Shared between	SNPs	INDELs
4	407,095	1,426
1	8,849,306	106,554
3	1,709,786	1,767
2	4,651,727	7,454
Total	15,617,914	117,201

Table 53. Base changes (SNPs) between CCSP bison and domestic cattle UMD3.1.

	A	C	G	T
A	0	626,415	2,676,252	526,706
C	658,731	0	585,673	2,734,383
G	2,743,063	588,643	0	662,212
T	525,950	2,671,340	626,356	0

Table 54. Transition and transversion counts and ratio for each CCSP sample and population totals found against domestic cattle reference UMD3.1.

Sample	48-5795	50-5792	61-5793	68-5784	Total
Transitions	8,535,821	9,888,021	7,555,401	7,710,099	33,689,342
Transversions	3,751,512	4,390,577	3,345,246	3,447,405	14,934,740
Ts/Tv	2.275	2.252	2.259	2.236	2.256

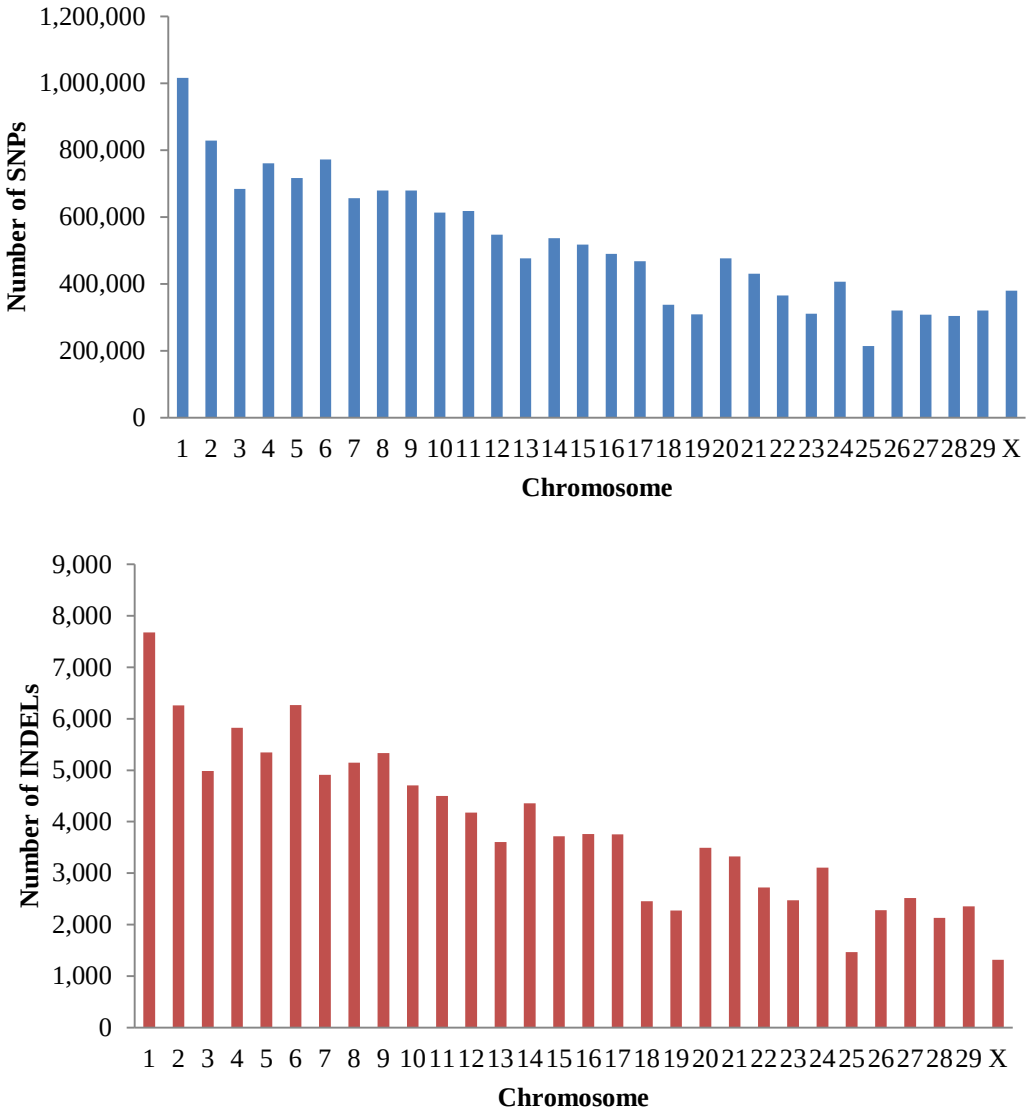


Figure 18. Number of SNPs (blue) and INDELs (red) by chromosome for CCSP bison mapped to domestic cattle UMD3.1.

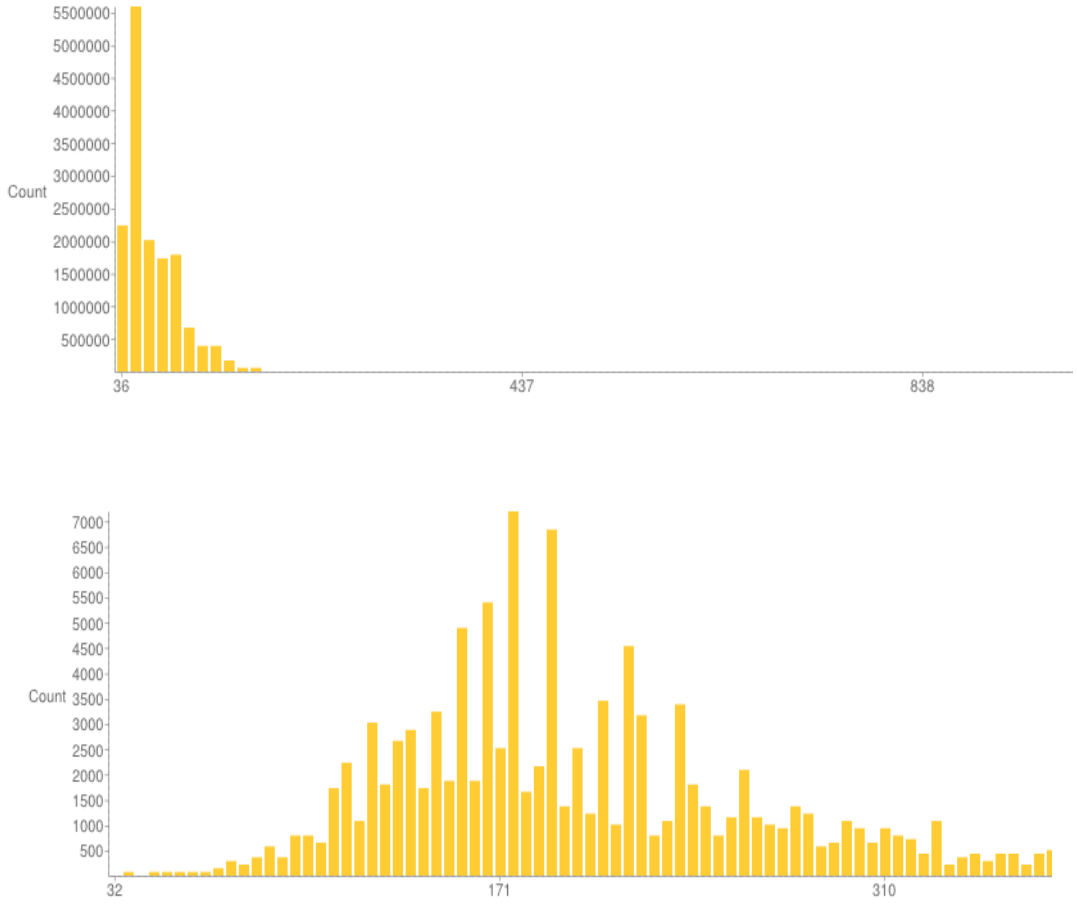


Figure 19. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for CCSP bison.

Summary of effects for SNPs and INDELs with SnpEff provides the number of effects by type and region (Table 55). For both SNPs and INDELs the majority of the variants were found in intergenic regions at 64.79% and 61.45%, respectively. Effects found for the SNPs and INDELs in CCSP (Appendices G and H). The biological functions of genes associated with the annotated SNPs and INDELs for CCSP where mainly protein coding genes for both variant types (Table 56). There were 24,506 genes

annotated from the CCSP SNPs and INDELs combined. Of those genes, 11,043 genes that were found only in the SNP gene file, and all of the genes annotated from the INDEL list were in the SNP gene list.

Table 55. Genomic regions associated with annotated SNPs and INDELs for combined CCSP samples.

	SNPs		INDELs	
Region Type (alphabetical order)	Count	Percent	Count	Percent
DOWNSTREAM	643,887	3.69%	5,869	4.38%
EXON	90,346	0.52%	122	0.09%
INTERGENIC	11,311,482	64.79%	82,304	61.45%
INTRON	4,630,911	26.52%	37,562	28.05%
NONE	85,566	0.49%	1,054	0.79%
SPLICE_SITE_ACCEPTOR	246	0.00%	27	0.02%
SPLICE_SITE_DONOR	246	0.00%	7	0.01%
SPLICE_SITE_REGION	9,374	0.05%	99	0.07%
TRANSCRIPT	26	0%		
UPSTREAM	651,438	3.73%	6,473	4.83%
UTR_3_PRIME	30,865	0.18%	376	0.28%
UTR_5_PRIME	4,995	0.03%	38	0.03%

Table 56. Biological functions of genes associated with annotated SNPs and INDELs in CCSP.

Biological Type	SNPs	INDELs
Protein coding	19,927	11,903
Ribosomal RNA	400	156
Miscellaneous RNA	175	71
Small nucleolar RNA	846	299
Pseudogene	622	227
Processed pseudogene	169	73
Micro RNA	1,150	324
Small nuclear RNA	1,217	410
Total	24,506	13,463

YNP

The raw Illumina paired-end sequences of the 4 YNP samples were first trimmed using FASTQ-MCF and bases with a quality score less than 20 from each individual read and minimum remaining sequence length of 70 (Aronesty 2011). They were then individually aligned to the domestic cattle UMD3.1 sequence using the same format as for the bison reference. The analysis for variants will remain separate for SNVs and INDELs and then grouped together to consider the variants within the population as a whole, the same as the results for the other bison populations.

Using the flagstat option in SAMtools statistics of the 4 YNP Illumina paired-end reads mapped to the domestic cattle reference sequence were obtained (Table 57). YNP1856 had the highest amount of reads to be aligned to the domestic cattle sequence and there was an average of 239,507,759 raw Illumina parried-end reads for the YNP bison to be compared to the domestic cattle reference sequence.

Table 57. Samtools flagstat statistics of the 4 YNP bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 (reads are in base pairs).

Statistic	YNP1856	YNP1861	2009005885	2009005899
in total (QC-passed reads + QC-failed reads)	321,846,448	276,808,942	190,921,092	168,454,552
duplicates	0	0	0	0
mapped	310,460,638 (96.46%)	267,227,578 (96.54%)	184,420,440 (96.6%)	163,357,168 (96.97%)
paired in sequencing	321,846,448	276,808,942	190,921,092	168,454,552
read1	160,923,224	138,404,471	95,460,546	84,227,276
read2	160,923,224	138,404,471	95,460,546	84,227,276
properly paired	299,322,318	257,284,042	178,114,274	158,026,685
with itself and mate mapped	306,460,745	263,591,367	182,401,073	161,610,113
singletons	3,999,893	3,636,211	2,019,367	1,747,055
with mate mapped to a different chr	6,486,703	5,685,503	3,771,208	3,191,585
with mate mapped to a different chr (mapQ>=5)	2,177,936	1,930,407	1,390,239	1,138,129

All of the samples were found to have similar amounts of SNPs and INDELs called, even with the varying amounts of sequence data to be analyzed for the analysis. Table 58 offers a summary for each individual sample and their SNVs detected, where the most SNVs detected were found for YNP1861 with 25,846,206 SNVs. The 4 YNP samples had more sequencing coverage than the EIW and CCSP samples, so the increased amount of variants seen for the YNP samples is expected.

Table 58. Individual variant summary statistics of 4 YNP found from comparing sequences to domestic cattle UMD3.1.

	YNP1856		YNP1861		2009005885		2009005899	
	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs	SNVs	INDELs
Homozygous Variant alleles	21,284,230	209,789	22,205,662	1,714,712	21,390,382	1,226,350	20,979,925	1,039,422
Heterozygous (one Reference one variantt)	3,825,192	24,624	3,624,586	152,847	3,036,493	76,778	2,839,584	57,829
Variant Count	25,128,463	234,970	25,846,206	1,869,902	24,436,194	1,303,736	23,826,893	1,097,620
Reference Alleles	3,825,192	24,624	3,624,586	152,847	3,036,493	76,778	2,839,584	57,829
Unique Alleles	1,264,467	32,803	2,111,114	512,418	1,221,876	172,212	1,031,702	117,915
Heterozygous Variant Alleles	19,041	557	15,958	2,343	9,319	609	7,384	369

Table 59. Summary statistics for SNPs and INDELs found in 4 YNP bison compared to domestic cattle.

	SNPs	INDELs
Warnings	3,075,414	239,955
Errors	118,806	4,036
Number of lines (input file)	30,462,746	2,303,504
Number of variants	30,538,894	2,332,245
Number of multi-allelic VCF entries	76,020	28,312
Number of effects	34,343,282	2,641,700
Genome total length	2,670,424,091	2,670,423,063
Genome effective length	2,660,908,197	2,660,907,169
Variant rate	1 variant every 87 bases	1 variant every 1,140 bases

SAMtools vcf-stats and SnpEff were used to analyze each individual VCF files and combined VCF file for the YNP population for statistics of SNVs and INDELs detected. A total of 30,538,894 total SNPs were annotated from the YNP bison and the domestic cattle reference sequence (Table 59). Of these total SNPs, 18,676,605 (61.1%) SNPs were found to be in common between all 4 YNP samples (Table 60). Overall there was one SNP detected within the YNP population every 87 bases with 34,343,282 effects detected within the genome (Table 59). The most common base change was a G>A transition with 5,499,684 (Table 61). The combined Ts/Tv raito for the YNP samples was 2.25, indicating that we were able to detect variants with very few or any false variants. Individual Ts/Tv ratios can be found in Table 62.

Table 60. Number of common variants found between the 4 YNP samples.

Shared between	SNPs	INDELs
4	18,676,605	65,951
1	3,872,417	835,348
3	4,831,471	602,666
2	3,082,253	799,539
Total	30,462,746	2,303,504

Table 61. Base changes (SNPs) between YNP bison and domestic cattle UMD3.1.

	A	C	G	T
A	0	1,217,846	5,091,039	993,430
C	1,279,119	0	1,203,655	5,475,854
G	5,499,684	1,207,596	0	1,284,417
T	992,130	5,069,904	1,224,220	0

Table 62. Transition and transversion counts and ratio for each YNP sample and population totals found against domestic cattle reference UMD3.1.

Sample	YNP1856	YNP1861	2009005885	2009005899	Total
Transitions	32,394,176	33,481,133	31,967,648	31,260,442	21,136,481
Transversions	14,037,558	14,586,693	13,868,247	13,553,760	9,402,413
Ts/Tv	2.31	2.3	2.31	2.31	2.25

There were 2,332,245 INDELs discovered between YNP bison and domestic cattle, with 1,101,381 (47.2%) insertions and 1,230,864 (52.8%) deletions. All INDELs were used for annotation and 1 INDEL was detected every 1,140 bases, with 2,641,700 genomic effects detected (Table 59). From the total detected INDELS there were 65,951 (2.86%) that were found in common of all 4 YNP bison samples. The number of variants found on each chromosome for both SNPs and INDELs for YNP aligned to domestic

cattle are depicted in Figure 20, with most of the variants found on Chromosome 1 for both type of variants. Figure 21 shows the count of quality scores for variants of those SNVs and INDELs annotated with SnpEff after filtering, respectively.

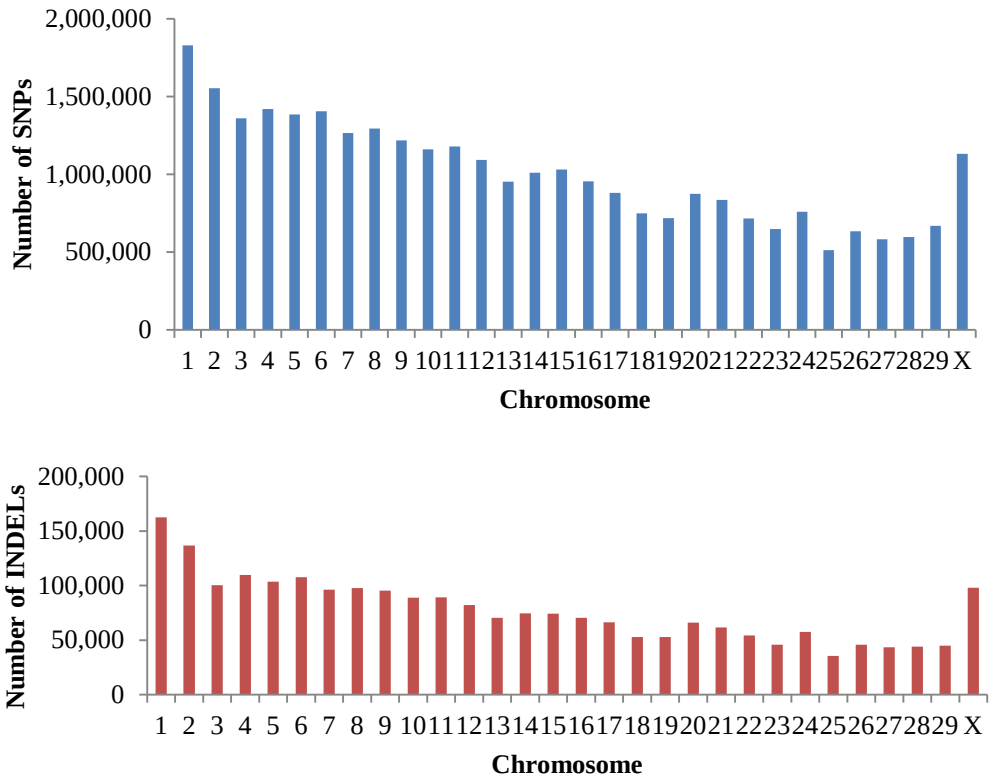


Figure 20. Number of SNPs (blue) and INDELs (red) by chromosome for YNP bison mapped to domestic cattle UMD3.1.

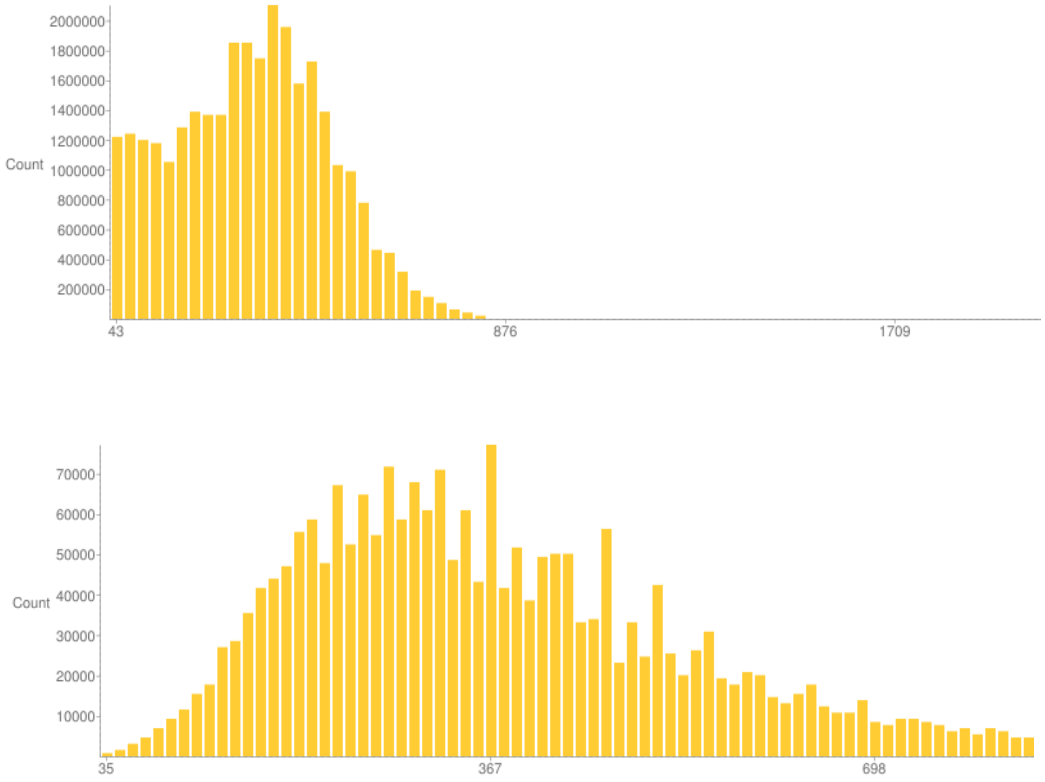


Figure 21. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) annotated for YNP bison.

Summary of effects for SNPs and INDELs with SnpEff provides the number of effects by type and region (Table 63). For both SNPs and INDELs the majority of the variants were found in intergenic regions and introns. There were 32 and 52 genomic effects that were found to be associated with the SNPs and INDELs annotated in YNP (Appendices I and J). The biological functions of genes associated with the annotated SNPs and INDELs for YNP were mainly protein coding genes for both variant types (Table 64). There were 24,524 genes annotated from the YNP SNPs and INDELs

combined. Of those total genes that were annotated, 116 genes were found to be in the SNP gene list and not in the INDEL gene list.

Table 63. Genomic regions associated with annotated SNPs and INDELs for combined YNP samples.

Region	SNPs		INDELs	
type (alphabetical order)	Count	Percent	Count	Percent
DOWNSTREAM	1,388,554	4.04%	118,102	4.47%
EXON	222,057	0.65%	2,459	0.09%
INTERGENIC	22,168,314	64.55%	1,658,440	62.78%
INTRON	8,976,113	26.14%	741,096	28.05%
NONE	118,806	0.35%	4,058	0.15%
SPLICE_SITE_ACCEPTOR	639	0.00%	238	0.01%
SPLICE_SITE_DONOR	601	0.00%	75	0.00%
SPLICE_SITE_REGION	21,921	0.06%	1,955	0.07%
TRANSCRIPT	74	0%	2	0%
UPSTREAM	1,364,841	3.97%	107,665	4.08%
UTR_3_PRIME	65,952	0.19%	6,892	0.26%
UTR_5_PRIME	15,410	0.05%	718	0.03%

Table 64. Biological functions of genes associated with annotated SNPs and INDELs in YNP.

Biological Type	SNPs	INDELs
Protein coding	19,938	19,847
Ribosomal RNA	400	391
Miscellaneous RNA	175	174
Small nucleolar RNA	846	845
Pseudogene	624	620
Processed pseudogene	171	168
Micro RNA	1,152	1,147
Small nuclear RNA	1,218	1,216
Total	24,524	24,408

Historic Samples

Historic sequences were trimmed using FASTQ-MCF and bases with a quality score less than 20 from each individual read and minimum remaining sequence length of 70 (Aronesty 2011), they were individually aligned to the domestic cattle UMD3.1 sequence for the purpose of identifying SNVs and INDELs separately by GATK. The historic samples were kept separated and then combined to evaluate the variants identified as a population, as the other samples were. The analysis for variants will remain separate for SNVs and INDELs.

The mem alignment option of BWA was used to align the filtered raw sequences of the historic bison samples using Illumina sequencing technology to the domestic cattle reference genome sequence (UMD3.1). The SAMtools options, view and flagstat, were used to obtain statistics of the historic bison Illumina paired-end reads mapped to the bison reference sequence. Historic sample S6 had more reads to align to UMD3.1 compared to S9 (Table 65). This could lead to more variants being detected for S6 that was seen in further analysis.

Historic sample S6 was found to have more SNVs and INDELs called, compared to S9 (Table 66). Table 46 offers a summary for each individual sample and their SNVs detected and INDELs found for the historic samples. Most of the SNVs for each individual were found to be homozygous for the variant allele, with the fewest SNVs that were heterozygous for 2 different variant alleles. S6 and S9 did have 27,509 and 8,045 SNVs that were heterozygous for two different variant alleles, respectively.

Table 65. Samtools flagstat statistics of the historic bison raw sequences mapped to the domestic cattle reference sequence UMD3.1 (reads are in base pairs).

Statistic	S6	S9
in total (QC-passed reads + QC-failed reads)	280,737,169	126,351,281
duplicates	0	0
mapped	279,692,711 (99.63%)	119,788,521 (94.81%)
paired in sequencing	280,737,169	126,351,281
read1	139,459,871	62,846,164
read2	141,277,298	63,505,117
properly paired	256,088,403	105,710,842
with itself and mate mapped	279,601,730	119,767,495
singletons	90,981	21,026
with mate mapped to a different chr	23,030,866	14,132,689
with mate mapped to a different chr (mapQ>=5)	9,895,935	6,576,562

Table 66. Individual summary statistics of 2 historic bison variants found from comparing sequences to domestic cattle UMD3.1.

	S6		S9	
	SNP	INDEL	SNVs	INDELs
Homozygous Variant alleles	15,909,197	689,724	3,877,017	73,459
Heterozygous (one Reference one variant)	8,991,312	149,487	4,921,122	63,096
Variant Count	24,928,018	840,450	8,806,184	136,682
Reference Alleles	8,991,312	149,487	4,921,122	63,096
Unique Alleles	21,816,979	840,450	5,695,145	91,389
Heterozygous Variant Alleles	27,509	1,239	8,045	127

SAMtools vcf-stats and SnpEff were used to analyze each individual VCF files for these historic samples for statistics of SNVs and INDELs detected and annotated (Table 67). For historic sample S6 there was 1 variant every 106 bases for SNVs and 1

variant every 3,161 bases for INDELs, each with 28,130,628 and 963,646 effects detected. There were 9,773,996 and 152,361 effects detected for SNVs and INDELs, respectfully for S9 and 1 SNV was detected every 301 bases while an INDEL was detected every 19,449 bases. The most common substitution between these historic samples and domestic cattle was a G>A substitution at 4,770,751 and 1,722,349 for S6 and S9, respectively, while the least common substitution is an A>T with 735,722 detected for S6 and A>T for S9 with only 302,319 detected (Table 68). The transitions to transversions ratio for S6 was calculated at 2.11, from 16,939,153 transitions and 8,016,374 transversions detected and there were 6,030,927 transitions and 2,783,302 transversions detected for S9, with a transition to transversion ratio of 2.17 (Table 68).

Table 67. Summary statistics for SNPs and INDELs found in 2 historical bison compared to domestic cattle UMD3.1.

	S6		S9	
	SNVs	INDELs	SNVs	INDELs
Warnings	2,495,754	90,589	721,813	11,616
Errors	169,488	3,166	128,699	1,232
Number of lines (input file)	24,928,018	840,450	8,806,184	136,682
Number of variants	24,955,527	841,688	8,814,229	136,808
Number of multi-allelic VCF entries	27,509	1,238	8,045	126
Number of effects	28,130,628	963,646	9,733,996	152,361
Genome total length	2,670,424,302	2,670,423,246	2,670,424,302	2,670,424,302
Genome effective length	2,660,908,408	2,660,907,352	2,660,908,408	2,660,908,408
Variant rate	1 variant every 106 bases	1 variant every 3,161 bases	1 variant every 301 bases	1 variant every 19,449 bases

Table 68. Base substitutions (SNPs) and transition and transversions found historic bison samples and domestic cattle UMD3.1.

Substitution	S6	S9
A>C	887,551	335,669
A>G	3,728,481	1,290,898
T>G	890,827	337,377
T>A	867,367	360,195
C>T	4,724,780	1,699,944
T>C	3,715,141	1,290,924
C>A	1,469,709	423,396
G>T	1,377,658	381,205
A>T	735,722	302,319
C>G	893,205	314,289
G>C	894,335	314,001
G>A	4,770,751	1,722,349
Transitions	16,939,153	6,030,927
Transversions	8,016,374	2,783,302
Ts/Tv ratio	2.11	2.17

There were 841,688 and 152,361 INDELs discovered for historic samples S6 and S9, respectively (Table 67). From the total INDELS there were 385,125 (47.76%) insertions and 456,563 (52.24%) deletions detected for S6, and 55,908 (40.9%) insertions and 80,900 (59.1%) deletions found for S9. All INDELs and SNVs were used for annotation in SnpEff, which gave chromosomal variant counts for both SNPs and INDELs for the historic bison samples aligned to domestic cattle are depicted in Figure 22, with most of the variants found on Chromosome 1 for both type of variants for both samples. Since these samples are older and contamination and degradation of DNA is an issue, the quality scores of each variant were considered and can be found in Figure 23.

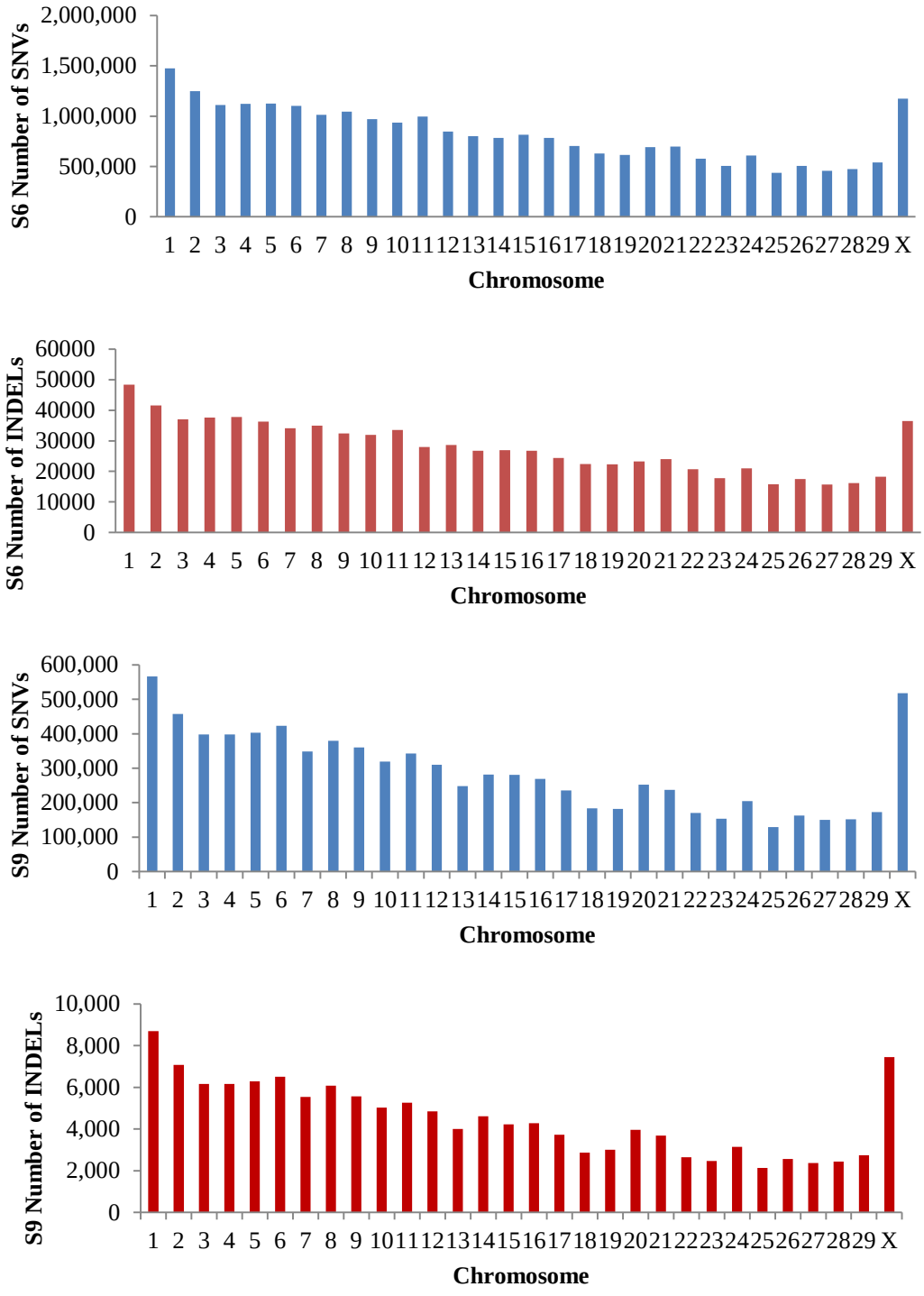


Figure 22. Number of SNVs (blue) and INDELs (red) by chromosome for historic bison samples mapped to domestic cattle UMD3.1 (S6 on top S9 on bottom).

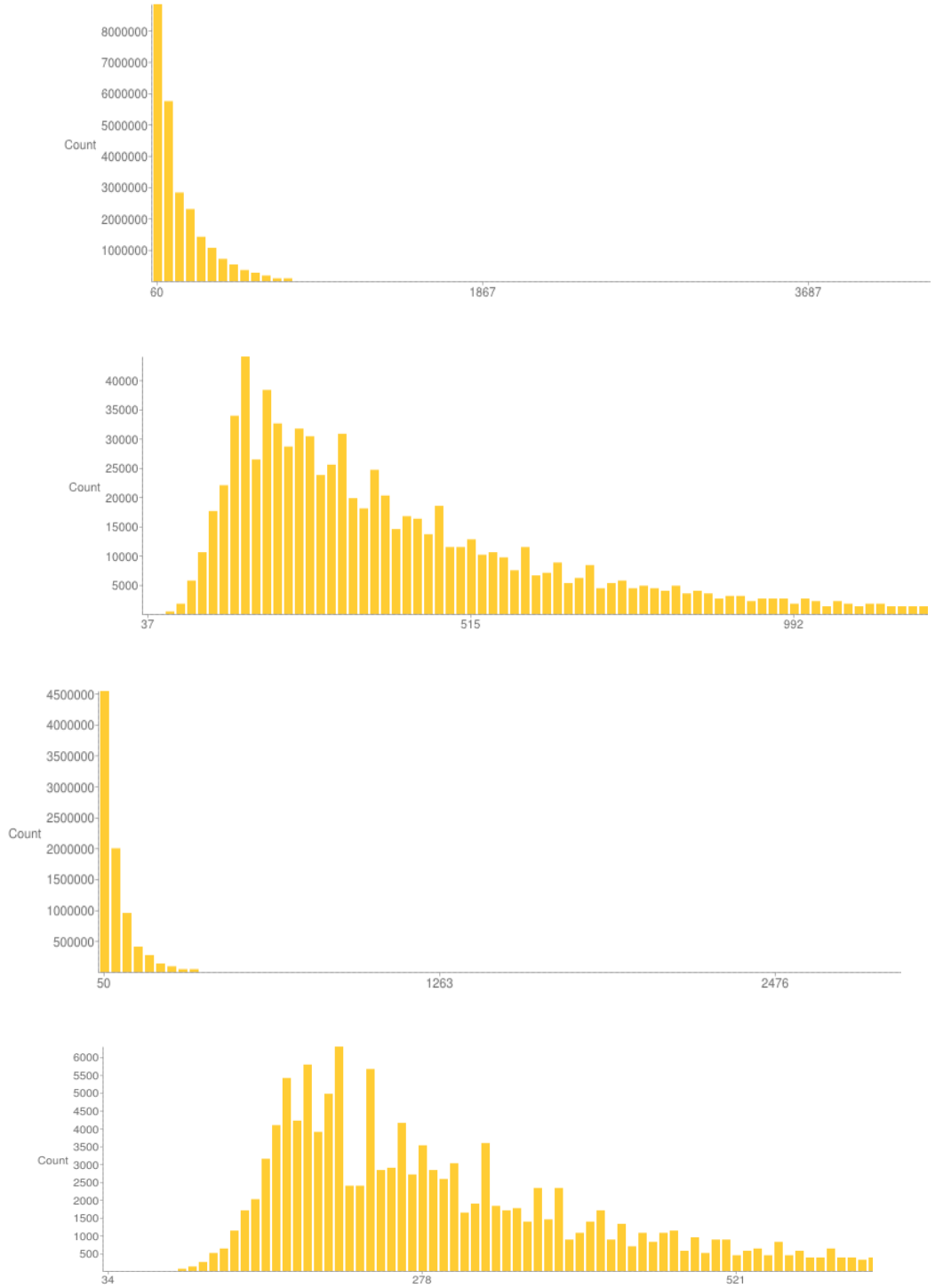


Figure 23. Variant count with corresponding quality scores of those SNPs (top) and INDELs (bottom) for historic bison samples (S6 top 2, S9 bottom 2).

Summary of effects for SNPs and INDELs with SnpEff provides the number of effects by type and region for both historical samples (Table 69). For both SNPs and INDELs the majority of the variants were found in intergenic regions for both samples. Individual effects found for the SNVs and INDELs for historic sample S6 and S9 can be found in Appendices K and L. The biological functions of genes associated with the annotated SNPs and INDELs for these historic bison samples where mainly protein coding genes for both variant types and corresponding numbers for each variant type and sample can be found in Table 70.

Table 69. Genomic region associated with annotated SNPs and INDELs for both historical bison, S6 and S9.

	S6				S9			
	SNVs		INDELs		SNVs		INDELs	
Region	Count	Percent	Count	Percent	Count	Percent	Count	Percent
DOWNSTREAM	1,157,949	4.12%	47,501	4.93%	309,780	3.18%	5,455	3.58%
EXON	223,433	0.79%	1,679	0.17%	12,187	0.13%	80	0.05%
INTERGENIC	17,997,844	63.98%	595,505	61.80%	6,669,655	68.52%	103,190	67.73%
INTRON	7,326,813	26.05%	267,210	27.73%	2,243,306	23.05%	36,141	23.72%
NONE	169,488	0.60%	3,181	0.33%	128,699	1.32%	1,232	0.81%
SPLICE_SITE_ACCEPTOR	894	0.00%	122	0.01%	66	0.00%	8	0.01%
SPLICE_SITE_DONOR	993	0.00%	52	0.01%	61	0.00%	2	0.00%
SPLICE_SITE_REGION	18,932	0.07%	732	0.08%	794	0.01%	12	0.01%
TRANSCRIPT	104	0%	1	0%	3	0%		
UPSTREAM	1,167,275	4.15%	44,735	4.64%	365,662	3.76%	6,175	4.05%
UTR_3_PRIME	52,540	0.19%	2,448	0.25%	3,260	0.03%	61	0.04%
UTR_5_PRIME	14,363	0.05%	480	0.05%	523	0.01%	5	0.00%

There were a total of 24,533 gene IDs found for historical sample S6 after combining the INDEL and SNV gene lists and removing duplicate Gene names and gene IDs. 93 of these genes were only found in the INDEL gene list and not the SNV list while only 749 of these genes were not in the INDEL gene list and 25,905 gene IDs were found in both SNV and INDEL gene lists.

Table 70. Biological functions of genes associated with annotated SNPs and INDELs in 2 historical samples, S6 and S9.

	S6		S9	
Biological Type	SNVs	INDELs	SNVs	INDELs
Protein coding	19,943	19,541	19,835	11,710
Ribosomal RNA	401	368	401	170
Miscellaneous RNA	175	164	175	60
Small nucleolar RNA	846	787	832	280
Pseudogene	626	581	624	222
Processed pseudogene	171	159	169	57
Micro RNA	1,152	1,068	1,103	310
Small nuclear RNA	1,219	1,135	1,216	458
Total	24,533	23,803	24,355	13,267

There were a total of 24,355 gene ID's annotated for historical sample S9 for both the SNVs and INDELs detected. There were 11,089 gene IDs that were in the SNV file but not in the INDEL file for historic sample S9. The gene lists were combined and ran through DAVID, as well as those SNV genes that were not found in INDELs gene list and the one single gene ID not found in the SNV gene file for S9. When comparing the historic samples gene lists, a total of 24,346 genes were found in common between

S6 and S9. With 187 S6 genes not being in found in S9, and 9 genes only being found in S9 and not S6 (Appendix M). Again, most of these 24,346 genes are associated with being protein-coding genes.

All 24,541 Gene IDs for the historical samples were analyzed in DAVID for a GO analysis, and 18,989 DAVID IDs were matched in the *Bos taurus* database. These results were the same from the DAVID analysis of Templeton to domestic cattle analysis in Chapter 1, which produced the same 48 GO terms with a FDR significant p-value ≤ 0.05 (Appendix A).

Comparison of All Bison Sample Variants and Annotated Genes to Domestic Cattle

All variant statistics and gene IDs were compared for EIW, CCSP, YNP, historic samples, and Templeton (bison references sequence; see Chapter 1), to give us a summary comparison of 15 bison samples to domestic cattle (Appendix O). Ignoring the samples that have less variant calls due to less sequence comparisons it is interesting to see that more heterozygous variant for two different variants were found for Templeton and the historic samples, S6 and S9, and YNP (Appendix O). This could be due to the fact that they had more reads considered for the analysis then the pooled samples. Therefore, comparing those Wood buffalo samples (26-1525 and 233-1676) that had good quality reads and CCSP samples, it is noticed that they have similar heterozygous variant for two variant allele's statistics for SNVs and INDELs.

With such varying sequences it is hard to do a real comparison unless within populations and do significance between population samples is detected. EIW, CCSP, and YNP samples were compared as populations, not as each individual sample for types

of SNP averages detected to domestic cattle (Table 71). As stated for each individual population, most of the SNVs are homozygous for one single variant. When comparing the numbers of reference alleles across populations, Templeton, YNP and the historical samples were found to have more reference alleles variants than CCSP and EIW populations. Again this could be due to less sequencing coverage for those populations which could have missed the parts of the sequences that contain heterozygous variants and caused homozygous calls to be inflated, as discussed previously.

Table 71. Comparison across average amounts of each populations or samples SNPs detected to domestic cattle reference UMD3.1.

	CCSP	EIW	YNP	Templeton	Historic Samples
	SNPs	SNPs	SNPs	SNPs	SNPs
Homozygous Variant alleles	0.9521	0.9562	0.6754	0.7761	0.8779
Heterozygous (one Reference one variant)	0.0478	0.0437	0.3236	0.2225	0.1217
Variant Count	15,617,914	9,590,819	30,538,894	28,443,364	48,263,087
Reference Alleles	0.0478	0.0437	0.3236	0.2225	0.1217
Unique Alleles	0.0213	0.0251	0.0488	0.0313	0.0194
Heterozygous Variant Alleles	0.0001	0.0001	0.0010	0.0014	0.0003

Table 72. Comparison of those homozygous SNPs found to be in common for historic samples and how many of those are in common to certain populations.

	Historic Samples	Percentage
Historic Samples	2,060,635	
Templeton	2,017,879	0.979
YNP	1,660,494	0.806
EIW	5,075	0.002
CCSP	68,110	0.033
Templeton and YNP	1,651,169	0.801
All	2,843	0.001

YNP is known to be the standard for bison genetics with no known detected hybrids found in the bison population. The genetics found at YNP are generally used for the standard to test other bison populations against. This allowed for comparing bison populations to what was believed to be bison without hybrids and if a bison population was found to have a new allele that was not previously found in YNP and EIW it was confirmed not to be of domestic cattle origin and in fact a unique allele for that population.

Using the historical samples SNPs that were homozygous (not a variant from the domestic cattle reference) that were detected when compared against the domestic cattle reference a VCF file was generated to compare all the remaining samples to. This generated a VCF file with 2,060,635 SNPs that were homozygous and in common between historical samples S6 and S9. This VCF was first compared to Templeton's VCF file containing the SNVs that were detected from the comparison to UMD3.1. Using VCF-isec to remove those SNVs that were only found to match the homozygous VCF file, a new VCF was produced with 2,017,879 (approximately 97.9%) SNPs to be homozygous, and in common with Templeton and the historical samples (Table 72).

The same method was done using just the historical samples homozygous SNPs VCF and the YNP samples. This produced 1,660,494 (80.6%) SNPs that were found to be homozygous and shared (informative) between the YNP and historical samples. The same was done for the EIW and CCSP samples, these had fewer SNPs in common with only 5,075 and 68,110 SNPs, respectively (Table 73). And when all samples were

compared only 2,843 SNPs were found to be homozygous and informative for bison to domestic cattle.

If YNP is the gold standard for bison and with added historical samples variants, the analysis of comparing those samples that were form YNP and historical samples was done. This allowed for a new VCF file that contained 1,651,169 (80.1%) of the 2,060,635 SNPs that were homozygous for the historical samples (Table 72). This allows to have over a million SNPs that are homozygous to the domestic cattle reference, that were determined to be informative for 6 bison samples who have YNP lineages.

All variants were summarized in Table 73, showing counts for SNVs, Insertions, and deletions, as well as the total variants found for CCSP, Wood buffalo populations and the individual sequences, Templeton, S6 and S9. There are a total of 50,746,586 variants found between these 15 bison and domestic cattle, with 47,514,082 SNPs, 1,492,303 insertions and 1,740,200 deletions. A SNP variant was detected every 56 bases, while and INDEL was detected every 823 bases.

Table 73. Variant Summary for EIW, CCSP, and YNP populations, as well as the individuals Templeton, S6 and S9.

Variant Type	CCSP	EIW	YNP	Templeton	S6	S9
SNP	15,617,914	9,590,819	30,538,894	22,073,944	24,955,527	8,806,184
INSERTION	55,773	22,994	1,101,381	1,233,140	385,125	162,921
DELETION	61,535	24,532	1,230,864	1,394,505	456,563	226,079
Total	15,735,222	9,638,345	32,871,139	24,701,589	25,797,215	9,195,184

All annotated gene lists were combined and analyzed for unique genes for populations. Duplicate gene IDs were removed if there was a duplicate gene name or transcript ID. There were a total of 24,559 genes annotated for all bison samples, for both SNPs and INDELs. A total of 24,286 (98.9%) genes were found to be in all of 15 bison samples. Historical sample S9 was found to not have 167 genes that were found in the other 14 bison samples (Appendix M). There were 2 genes that were not found in Templeton, but the other 14 samples, and also 2 genes that were annotated in only Templeton and not the other samples (Appendix P).

There were other unique circumstances for the complete gene list. Most of these were when one population was not found to have variants annotated for that gene. For example, there were 17 genes that were found for all of the bison samples, except the EIW population and 12 genes that were not found to be in the CCSP samples (Appendix P). There were 3 genes that were found to not be in both historical samples, while only 2 genes were annotated from variants from only S9. Lastly, there were 16 genes that were annotated from variants from YNP, Templeton and the historical samples (Table 74). These 16 genes could be specific to only YNP bison since they were found only in those bison that were from YNP and also the area that would become YNP (historical samples). This will require future investigation to verify these genes could only be found in bison with YNP lineage and what the function is in these bison and why other bison populations might not have them.

Table 74. Unique annotated genes for YNP (Y), Templeton (T), and historic samples (O) after comparison of all bison gene lists found from annotation of variants identified to domestic cattle.

Gene Name	Gene ID	Transcript ID	Biological Type	Notes
BOLA-DQB	ENSBTAG00000021077	ENSBTAT00000046279	protein_coding	YTO not WC
ENSBTAG00000001291	ENSBTAG00000001291	ENSBTAT00000055982	protein_coding	YTO not WC
ENSBTAG00000001383	ENSBTAG00000001383	ENSBTAT00000001819	pseudogene	YTO not WC
ENSBTAG00000026078	ENSBTAG00000026078	ENSBTAT00000027976	protein_coding	YTO not WC
ENSBTAG00000030925	ENSBTAG00000030925	ENSBTAT00000043770	pseudogene	YTO not WC
ENSBTAG00000033055	ENSBTAG00000033055	ENSBTAT00000046929	protein_coding	YTO not WC
ENSBTAG00000033445	ENSBTAG00000033445	ENSBTAT00000047574	protein_coding	YTO not WC
ENSBTAG00000035988	ENSBTAG00000035988	ENSBTAT00000004342	protein_coding	YTO not WC
ENSBTAG00000037875	ENSBTAG00000037875	ENSBTAT00000057015	protein_coding	YTO not WC
ENSBTAG00000045675	ENSBTAG00000045675	ENSBTAT00000057342	protein_coding	YTO not WC
ENSBTAG00000046884	ENSBTAG00000046884	ENSBTAT00000065415	protein_coding	YTO not WC
ENSBTAG00000047041	ENSBTAG00000047041	ENSBTAT00000064286	protein_coding	YTO not WC
ENSBTAG00000047308	ENSBTAG00000047308	ENSBTAT00000064592	miRNA	YTO not WC
ENSBTAG00000047360	ENSBTAG00000047360	ENSBTAT00000052551	pseudogene	YTO not WC
U6	ENSBTAG00000046479	ENSBTAT00000065172	snRNA	YTO not WC
U6	ENSBTAG00000047456	ENSBTAT00000064526	snRNA	YTO not WC

Some other unique annotated genes that were found in certain situations, including those that were genes found in CCSP (C) and wood buffalo (W) populations and Templeton (T), but not in YNP (Y) and the historical samples (O); and those that were found in Templeton and historical samples, and not EIW, CCSP or YNP populations (Table 75).

With these different situations from comparing all of the bison samples, there were 40 annotated genes found to be unique to certain populations. These genes found to be unique in varying circumstances should be considered for future research to see

what the function could be in different bison populations or if other bison populations could have the same annotated genes from similar variants to domestic cattle.

Table 75. Unique genes after comparison of all bison genes found from annotation of variants identified to domestic cattle (see the notes section; abbreviations C=CCSP, Y=YNP, T=Templeton, O=historical samples and W=EIW).

Gene Name	Gene ID	Transcript ID	Biological Type	Notes
ENSBTAG00000047653	ENSBTAG00000047653	ENSBTAT00000065090	protein_coding	in CT not WYO
ENSBTAG00000048089	ENSBTAG00000048089	ENSBTAT00000064841	protein_coding	in CT not WYO
ENSBTAG00000039444	ENSBTAG00000039444	ENSBTAT00000057402	protein_coding	in CYTS6 not WS9
ENSBTAG00000046993	ENSBTAG00000046993	ENSBTAT00000034439	protein_coding	in CYTS6 not WS9
RAD51AP1	ENSBTAG00000040065	ENSBTAT00000012163	protein_coding	in CYTS6 not WS9
TRAV18	ENSBTAG00000045863	ENSBTAT00000064546	protein_coding	in CYTS6 not WS9
5S_rRNA	ENSBTAG00000028498	ENSBTAT00000040879	rRNA	in TO not in WCY
ENSBTAG00000008177	ENSBTAG00000008177	ENSBTAT00000010751	pseudogene	in TO not in WCY
ENSBTAG00000047292	ENSBTAG00000047292	ENSBTAT00000063761	protein_coding	in TO not in WCY
ENSBTAG00000035059	ENSBTAG00000035059	ENSBTAT00000049535	protein_coding	in TS6 not in WCYS9
ENSBTAG00000046442	ENSBTAG00000046442	ENSBTAT00000040381	protein_coding	in TS6 not in WCYS9
ENSBTAG00000033558	ENSBTAG00000033558	ENSBTAT00000049902	protein_coding	in WCT not in YO
ENSBTAG00000034761	ENSBTAG00000034761	ENSBTAT00000049218	protein_coding	in WCT not in YO
ENSBTAG00000040370	ENSBTAG00000040370	ENSBTAT00000065402	protein_coding	in WCT not in YO
ENSBTAG00000045738	ENSBTAG00000045738	ENSBTAT00000066049	protein_coding	in WCT not in YO
ENSBTAG00000046657	ENSBTAG00000046657	ENSBTAT00000064117	protein_coding	in WCT not in YO
PRAME	ENSBTAG00000046669	ENSBTAT00000050480	protein_coding	in WCT not in YO
PRAME	ENSBTAG00000047085	ENSBTAT00000063108	protein_coding	in WCT not in YO
TSPY-M2	ENSBTAG00000031517	ENSBTAT00000044671	protein_coding	in WCT not in YO
TSPY-M2	ENSBTAG00000048051	ENSBTAT00000048960	protein_coding	in WCT not in YO
ENSBTAG00000035012	ENSBTAG00000035012	ENSBTAT00000049474	protein_coding	in WCTS6 not in YS9
ENSBTAG00000033620	ENSBTAG00000033620	ENSBTAT00000047761	protein_coding	in WCTS9 not in YS6
HSFY2	ENSBTAG00000046306	ENSBTAT00000051978	protein_coding	in WCTS9 not in YS6
U6	ENSBTAG00000046483	ENSBTAT00000064707	snRNA	in WCTS9 not in YS6
ENSBTAG00000048193	ENSBTAG00000048193	ENSBTAT00000064582	protein_coding	in WYO not CT
ENSBTAG00000040382	ENSBTAG00000040382	ENSBTAT00000057394	protein_coding	in WYTS6 not CS9
ENSBTAG00000045777	ENSBTAG00000045777	ENSBTAT00000065421	protein_coding	in WYTS6 not CS9
ENSBTAG00000046595	ENSBTAG00000046595	ENSBTAT00000063976	protein_coding	in WYTS6 not CS9
ENSBTAG00000047817	ENSBTAG00000047817	ENSBTAT00000066183	protein_coding	in WYTS6 not CS9
ENSBTAG00000032833	ENSBTAG00000032833	ENSBTAT00000053697	protein_coding	in WYTS6 not CS9
MGC140151	ENSBTAG00000047295	ENSBTAT00000065824	protein_coding	YNP not WCTO
ENSBTAG00000046527	ENSBTAG00000046527	ENSBTAT00000064135	protein_coding	YS9 not WCTS6
OR8I2	ENSBTAG00000046285	ENSBTAT00000052309	protein_coding	YS9 not WCTS6
ENSBTAG00000003769	ENSBTAG00000003769	ENSBTAT00000004901	processed_pseudogene	YTS6 not WCS9
ENSBTAG00000045807	ENSBTAG00000045807	ENSBTAT00000063213	protein_coding	YTS6 not WCS9
ENSBTAG00000047051	ENSBTAG00000047051	ENSBTAT00000065262	processed_pseudogene	YTS6 not WCS9
ENSBTAG00000047404	ENSBTAG00000047404	ENSBTAT00000028788	protein_coding	YTS6 not WCS9
ENSBTAG00000047804	ENSBTAG00000047804	ENSBTAT00000063367	protein_coding	YTS6 not WCS9
ENSBTAG00000048199	ENSBTAG00000048199	ENSBTAT00000063562	protein_coding	YTS6 not WCS9
U6	ENSBTAG00000045512	ENSBTAT00000062983	snRNA	YTS6 not WCS9

DAVID analysis was done using just the combined gene list of 24,559 genes and 18,992 DAVID ID's were matched in the *Bos taurus* database. These results were similar to that of the historical samples and Templeton DAVID analysis, producing the same 48 GO terms that had a significant FDR value. To investigate those 167 genes that were not identified in historical sample S9 a DAVID analysis was also done on these genes. This produced a match of 118 DAVID IDs and only 12 GO terms that were found to have a significant p-value ≤ 0.05 (Table 76). Of these 12 GO terms, 6 were not found to be in the GO analysis from the combined all gene list.

Table 76. 12 significant GO terms resultant from gene list of those genes not annotated in S9 and then analyzed in DAVID for a GO analysis with the biological (GO) term, p-value and false discovery rater (FDR). Only terms with p-value of ≤ 0.05 were reported.

Biological Term	P-Value ≤	FDR	Notes
GO:0030900~forebrain development	0.0009	1.2075	
GO:0045165~cell fate commitment	0.0012	1.5538	not found in All gene list DAVID analysis
GO:0003002~regionalization	0.0020	2.6186	
GO:0007389~pattern specification process	0.0046	5.8346	
GO:0045449~regulation of transcription	0.0052	6.5943	not found in All gene list DAVID analysis
GO:0006355~regulation of transcription, DNA-dependent	0.0075	9.4631	not found in All gene list DAVID analysis
GO:0051252~regulation of RNA metabolic process	0.0083	10.4518	not found in All gene list DAVID analysis
GO:0001501~skeletal system development	0.0085	10.5811	
GO:0030182~neuron differentiation	0.0086	10.8117	
GO:0009952~anterior/posterior pattern formation	0.0154	18.4733	
GO:0021983~pituitary gland development	0.0416	42.8247	not found in All gene list DAVID analysis
GO:0021536~diencephalon development	0.0483	47.9153	not found in All gene list DAVID analysis

Phylogenetic Tree

Using the combined VCF file to domestic cattle for all 15 bison samples, SNPhylo was used to generate a phylogenetic tree (Figure 24). Based on this tree, the historic samples were placed next to the Yellowstone samples, with the YNP samples first and then Templeton and the CCSP samples were also placed closely next to the historic samples. A split between EIW and the historic samples was seen since the EIW samples were placed next to the YNP samples, which with the moving of YNP bison into EIW, this could be expected. What was also expected to be seen is the split between the CCSP and EIW populations. Since these are three different sub-populations being compared, it is interesting to point out that the samples that are representing southern plains bison are the farthest bison population from the EIW population.

What was not expected was to see a split between one of the CCSP sample from the others. The placement of CCSP 50-5792 within the EIW samples will require further analysis evaluate this split of the CCSP bison samples.

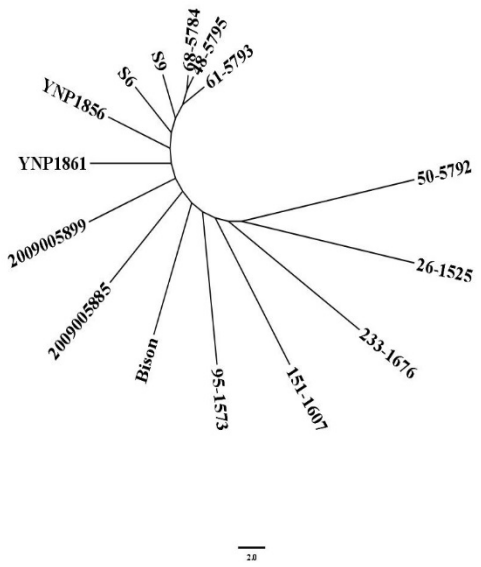


Figure 24. Phylogenetic tree from the combined VCF file for all 15 bison samples to UMD3.1 (Bison=Templeton).

Discussion

With the number of genomic variants detected across all 15 samples throughout the bison genome, an increase of variants has been found to help distinguish bison from domestic cattle at the genomic level, which was not previously known. These variants offer a more in depth analysis of the bison genome to determine introgression of domestic cattle genetics compared to what current technologies offer.

Annotated genes were identified from these variants to be used for gene comparisons across all 15 samples. These annotated genes can be used for future

analysis to better understand why bison and domestic cattle react differently to environment or diseases. In addition, by directly comparing whole genomes between domestic cattle and bison reference sequences, it is possible to define individual differences between these two closely related species and provide a foundation for developing an extremely robust test for introgression in bison. We have also identified informative SNPs between different bison populations and the historical samples. Primarily, we identified approximately 1,651,169 informative SNPs that needed to be validated in the future, which can then be used to define a more robust introgression test for bison. This analysis provides us with the first genomic level analysis of detected genomic variants and their associated genes from comparing bison genomic sequences to the domestic cattle reference sequence to detect introgression more thoroughly.

The phylogenetic tree that was used to determine taxonomy of bison variants to domestic cattle for all 15 bison samples suggests that further evaluation of the bison at EIW and CCSP needs to be done in the future. With most of the tree produced confirming prior expectations of the bison samples used, the different placements of CCSP with the EIW bison was intriguing to see. The placement of these bison within different populations on the phylogenetic tree shows that the CCSP and EIW bison could have had outside genomic influence from different bison herds and even from each other, this again will need future evaluation.

Conclusions

With the new genomic variants found across all 15 bison and domestic cattle we have vastly increased the number of variants that define the genomic differences between bison and domestic cattle. In total, 24,497 genes were annotated from these variants, with the majority of the annotated genes being protein coding genes. Genomic variants between re-sequenced bison samples and Templeton allowed for a multi-way comparison of bison genome sequences identifying unique variants for each population that were not previously known.

A summary of the variants detected for the bison populations and individuals Templeton, S6 and S9 can be found in Table 77. This makes comparison of the 15 bison samples variants detected to UMD3.1 (domestic cattle) and also UMD1.0 (bison). It is evident that there were more variants detected for all bison populations to UMD3.1. Templeton was only compared to UMD3.1 since he was the bison reference (UMD1.0).

To compare the types of SNPs detected for the bison samples to UMD3.1 and UMD1.0, the data was averaged for CCSP, EIW, and YNP and the historical samples (OS) into Table 78. Templeton again was left out of the comparison to himself. The point of interest for doing this was to show that CCSP and EIW were found to have on average less reference alleles to UMD1.0 and more homozygous variants, compared to YNP, Templeton, and OS. The history of these herds could lead to this outcome, since they are believed to be different classifications of sub-species of bison. This could also be due to the lower coverage of sequencing done for these samples

Table 77. Total variants found for bison populations and individuals to both UMD3.1 and UMD1.0

	CCSP		EIW		YNP		Templeton	S6		S9	
Variant Type	UMD3.1	UMD1.0	UMD3.1	UMD1.0	UMD3.1	UMD1.0	UMD3.1	UMD3.1	UMD1.0	UMD3.1	UMD1.0
SNP	15,617,914	3,877,737	9,590,819	2,192,618	30,538,894	9,157,950	22,073,944	24,955,527	11,857,832	16,951,692	6,635,219
INS	55,773	14,769	22,994	6,408	1,101,381	208,771	1,233,140	385,125	112,949	162,921	35,791
DEL	61,535	13,683	24,532	5,593	1,230,864	202,350	1,394,505	456,563	134,501	226,079	49,406
Total	15,735,222	3,906,189	9,638,345	2,204,619	32,871,139	9,569,071	24,701,589	25,797,215	12,105,282	17,340,692	6,720,416

Table 78. Averages of SNPs idenfitifed for all of the 15 bison samples to UMD3.1 and UMD1.0.

	CCSP		EIW		YNP		Templeton	OS	
Types of SNPs	UMD3.1	UMD1.0	UMD3.1	UMD1.0	UMD3.1	UMD1.0	UMD3.1	UMD3.1	UMD1.0
Homozygous Variant alleles	0.9521	0.7730	0.9562	0.7544	0.6754	0.4376	0.7761	0.8779	0.2370
Heterozygous (one Reference one variant)	0.0478	0.2267	0.0437	0.2451	0.3236	0.5621	0.2225	0.1217	0.7626
Variant Count	15,617,914	3,877,737	9,590,819	2,192,618	30,538,894	22,073,944	28,443,364	30,623,163	17,756,669
Reference Alleles	0.0478	0.2267	0.0437	0.2451	0.3236	0.5621	0.2225	0.1217	0.7626
Unique Alleles	0.0213	0.3718	0.0251	0.4020	0.0488	0.3073	0.0313	0.0194	0.9138
Heterozygous Variant Alleles	0.0001	0.0003	0.0001	0.0005	0.0010	0.0003	0.0014	0.0003	0.0005

All of the data produced from the comparison of the bison samples to the domestic cattle and bison reference sequences will act as a baseline for future comparisons of variants to these references. The data will be placed into a database that can be searchable for future research. With the addition of future bison sequences for comparison to both of these reference sequences, will allow for the validation of the data from this study and also provide new data to be added to this database.

Future studies will be needed to answer some of the questions that the data from this study has produced. For instance, the CCSP and EIW samples placement on the phylogenetic tree might confirm that wood buffalo and CCSP (as a representative of southern plains bison) could be valid sub-species since at the genomic level they can be placed within on different ends on a phylogenetic tree, but also could have influence from CCSP to EIW due to movement of animals into EIW from CCSP. The first step would need to validate the information of these samples from Elk Island to ensure these samples are in fact wood buffalo based on phenotypic characteristics and not plains bison that were introduced into the herd. This could show that with the introgression of plains bison into the wood buffalo population at Elk Island, the genomic integrity of the wood buffalo population has changed to represent plains bison for certain individuals.

For CCSP the data suggests that most of the variants detected to both UMD3.1 and UMD1.0 are primarily homozygous variants and also had the least detected reference alleles to UMD1.0. CCSP has been managed as a closed herd until 2004 when Halbert *et al.* determined the low levels of heterozygosity of this herd was putting the future of this herd in peril. Outside genetics were brought in to help increase the levels

of heterozygosity, but the samples used from CCSP were born before the introduction and use of these outside bulls. The low levels of heterozygosity is evident with the higher levels of homozygous variants detected for CCSP to both UMD3.1 and UMD1.0 and the genomic levels of heterozygosity and homozygosity need to be evaluated for these samples to see how they compare to the levels of the herd found by Halbert *et al.* 2004 using microsatellite data. However, to ensure that these low levels of heterozygosity is not due to over-estimation of homozygous variants detected for CCSP, as well as EIW, future sequencing at a deeper coverage is needed to validate the homozygosity level of these herds.

Using historical samples for re-sequencing allowed for the analysis of comparison of pre-population bottleneck bison to modern bison. Even with being able to detect approximately 12 and 9 million genomic variants between the historic samples S6 and S9, respectively, and Templeton, the percent of the genome that has detected variants is only 0.43 and 0.24%. Instead of focusing on what part of the genome has been lost or changed further analysis would need to focus on the annotated genes from these variants and what they could be controlling and influencing. For further analysis of the variants detected for the historic samples, heterozygosity of the samples could be ran for all of the 15 bison samples and compared to the historic samples to evaluate how the diversity has changed between the historic samples to the different populations.

Lastly, using a stricter criterion for analyzing homozygous variants of the historic samples, that were also informative between the two, we were able to make a dataset that was determined to be the baseline of genomic variants between bison and domestic

cattle. With comparing this dataset to YNP samples and Templeton we were able to identify over 1.6 million informative SNPs between all 7 individuals. With using this new dataset we can validate it by future re-sequenced bison samples to make a more robust test for introgression with bison into domestic cattle.

As with most research, this study has provided an abundant amount of information that can be used for future studies in bison and poses more questions to be answered. Whole genome sequencing has allowed us to greatly increase the genomic variant information between bison and domestic cattle to analyze where bison and domestic cattle are different in the genome. We have provided a foundation of data sets that can be built upon by future studies and further re-sequencing of bison samples. Using this foundation of data and analyses between bison and domestic cattle and within bison populations we can now better understand bison at the genomic level.

CHAPTER IV

CONCLUSIONS AND FUTURE RESEARCH

The history of the North American bison is one that can be seen as one of the few wildlife population recovery success stories, from having a total population reduced by almost 99% and a handful of survivors to now numbering approximately around 500,000. With the completion of the 2.82-Gb *de novo* reference assembly of the American bison genome, bison genetic research has now advanced into the genomic technology era. With the annotation of the bison *de novo* reference genome we were able to identify a total of 26,001 genes and pseudogenes with 20,782 genes being protein coding genes for bison. These new genes can be utilized in future research to examine how bison function at the genetic level for production traits, disease, and other phenotypic traits.

Most bison in North America today are in production herds, and bison meat has become a luxury protein source that is leaner and healthier than beef. Most bison producers are intrigued as to why bison have this ability to be this premium meat source. Using the variants detected and their associated annotated genes, future research can be done to evaluate which genes could be responsible for these production traits in bison, using a similar model that has previously been done by researchers over the past 20 years after the completion of the cattle reference genome.

With the new genomic variants found across all 15 bison and domestic cattle we have vastly increased the number of variants that define the genomic differences between bison and domestic cattle. In total, 24,497 genes were annotated from these

variants, with the majority of the annotated genes being protein coding genes. Future research can be done to evaluate how bison and domestic cattle react differently to vaccines for diseases to help combat diseases in bison, such as tuberculosis and brucellosis. Vaccine trials could be done to compare how annotated genes from variants for bison associated with disease or immunity could affect how bison react to vaccines or response to diseases. With the brucellosis and tuberculosis status of some bison populations these studies are imperative for bison populations to ensure their health for future generations.

Using the foundation dataset of approximately 1.6 million informative SNPs between bison and domestic cattle, a SNPChip can be generated to better test for introgression of domestic cattle genes into bison. The first step in doing this would be to validate this foundation dataset with future re-sequencing of bison to achieve allelic frequencies of the bison variants found to domestic cattle and also include any new variants that had not been previously detected from different bison populations. The allelic frequencies of the bison variants would also add a statistical power to this test, and would show how common certain alleles were found and which alleles could be rare, implying they are found in only certain populations or rare.

This bison SNPChip would also need the domestic cattle variants that have been found in the bison genome that are not due to ancestral influence. The best way to do evaluate this would be to re-sequence known hybrids of bison and domestic cattle crosses. This would be done with F1s, F2s, F3s, and so forth to compare the generational influence of domestic cattle genes into bison. Hybrid crosses that occurred

in past generations should also be evaluated but might be harder to determine how far back the cross occurred without records. With prior knowledge and historical records, we know which breeds of cattle were bred with bison to make hybrids, and using this information we could determine the variants for domestic cattle into bison at the positions of the foundation SNP dataset. With the appropriate amount of samples we could also determine allelic frequencies as would be done with bison to know which cattle alleles occur more frequently in bison and which alleles are rarer. This would allow us to have not only bison alleles, but known cattle alleles that could be within hybrid bison genomes to test unknown bison samples across the SNPChip and add a statistical power to the introgression test, making it more robust.

Ancestral parts of the genome could be determined between bison and domestic cattle using the ABBA-BABA test, and to also determine when introgression of domestic cattle could occur in the bison genome. By using an outlier species, such as river or water buffalo, the historic samples, a domestic cattle breed, and Templeton we can determine the D-statistic to analyze where all of the samples would be placed on the ancestral tree. This test could then be re-ran using the EIW, CCSP, and YNP samples. This would be of interest to do comparing the individual EIW samples since they were placed throughout the phylogenetic tree and not as one branch. Determining the parts of the genome that could be due to their common ancestor would be of interest to determine if these ancestral alleles have been conserved over time from the historical samples to modern samples, and what genes they could be influencing for bison and domestic cattle.

Genomic variants between re-sequenced bison samples and Templeton allowed for a multi-way comparison of bison genome sequences identifying unique variants for each population that were not previously known. Since the population bottleneck of bison the question seems to be how this has changed the diversity of the bison genome. Using the variant files generated for either the comparison to bison or domestic cattle, we could look at the heterozygosity values of the 15 bison samples. The main comparison would be the other 13 bison samples to the historical samples. This analysis would also need to take into consideration the management practices of the herds the bison came from. Knowing the poor genetic health of the CCSP bison herd that was determined by Halbert *et al.* in 2004 and the herd being closed from outside animals until 2004 could play more of a recent influence on the genomic diversity of this herd than the population bottleneck in the 1880's.

A future research project to better evaluate the genomic diversity of not only these herds, but other bison populations would be to re-sequence animals that occurred in these herds before the bottleneck, or shortly after. This would compare the genomic diversity of the bison around the time of the bottleneck and without any recent breeding management influences. Another way to study the effect of the bottleneck on genomic diversity would be to a decade comparison of samples for either one population or many, to see if there is a trend in genomic diversity with the increase of bison numbers over the years. This could show the genomic diversity of bison pre-bottleneck, during bottleneck, right after bottleneck, and during the recovery of bison over the years to current bison genomic levels.

The taxonomy question of wood buffalo and plains bison, was only complicated with the phylogenetic tree that was generated from the SNP data. The EIW samples were placed in different positions on the tree and not grouped together or even placed on the same branch. Information is needed on the samples that were sent from Elk Island to ensure these were in fact wood buffalo samples. The best way to evaluate the genome of wood buffalo would be to re-sequence historical samples of wood buffalo and then compare to Templeton and also the EIW and historical samples. We could evaluate the genomic differences of the historic plains bison and wood buffalo and how much of the genome is different between the two. Then compare the old wood buffalo samples to Templeton and EIW and determine the genomic differences that exist today between modern plains bison, and modern wood buffalo samples. From this we could determine if certain EIW samples have more plains bison genomics instead of wood buffalo genomics to help explain the placement of the EIW samples on the phylogenetic tree.

The same study could be done using the CCSP herd and a historical sample from this herd to see if they contain unique alleles that could distinguish them as southern plains bison. This could be done by doing another multi-way comparison of historic samples, modern plains bison, and the CCSP bison today. This would hopefully determine if the CCSP have variants to Templeton that could be unique only to that herd because of they are are descendants of southern plains bison.

Lastly, for bison conservation more bison populations need to be evaluated to determine the genomic importance of different bison herds. The most important part of the conservation stand point would be to find if certain herds have higher influence of

domestic cattle genomics that have occurred more recently than after the population bottleneck. The next important reason to evaluate more populations would be to determine which herds have better genetic diversity than others and if the herd could be in danger of low levels of genetic diversity at the genomic level. Finding herds that could have unique genomics could be used to establish satellite herds and help create new conservation herds. This could be helpful if certain herds have reached their carrying capacity of the land and need to be relocated.

This analysis provides the first genomic level analysis of detected genomic variants and their associated genes from comparing bison genomic sequences to the domestic cattle, which can be used to define a more robust introgression test for bison. Whole genome sequencing has provided a genomic foundation that can go more in depth for bison conservation and management that can define how bison have been able to thrive and survive over the years after a devastating population bottleneck that decreased their numbers greatly.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang et al., 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

Aronesty, E., 2011 ea-utils: Command-line tools for processing biological sequencing data; https://code.google.com/p/ea-utils/

Bosse, M., H.-J. Megens, L. A. Frantz, O. Madsen, G. Larson et al., 2014 Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature communications 5.

Boyd, D. P., 2003 Conservation of North American bison: status and recommendations, MS Thesis. University of Calgary, Alberta, Canada.

Buntjer, J., M. Otsen, I. Nijman, M. Kuiper and J. Lenstra, 2002 Phylogeny of bovine species based on AFLP fingerprinting. Heredity 88: 46-51.

Cai, L., J. F. Taylor, R. A. Wing, D. S. Gallagher, S.-S. Woo et al., 1995 Construction and characterization of bovine bacterial artificial chromosome library. Genomics 29: 413-425.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos et al., 2009 BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen et al., 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6: 80-92.

Coder, G. D., 1975 The national movement to preserve the American buffalo in the United States and Canada between 1880 and 1920. Ph.D. Dissertation (History). The Ohio State University, Columbus, OH.

Cronin, M. A., M. D. MacNeil, N. Vu, V. Leesburg, H. D. Blackburn et al., 2013 Genetic variation and differentiation of bison (*Bison bison*) subspecies and cattle (*Bos taurus*) breeds and subspecies. Journal of Heredity 104: 500-509.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks et al., 2011 The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

Dary, D., 1989 The buffalo book: the full saga of the American animal. Swallow Press/Ohio University Press, Chicago, IL.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire et al., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491-498.

Derr, J. N., P. W. Hedrick, N. D. Halbert, L. Plough, L. K. Dobson et al., 2012 Phenotypic effects of cattle mitochondrial DNA in American bison. Conservation Biology 26: 1130-1136.

Di Meo, G. P., A. Perucatti, S. Floriot, D. Incarnato, R. Rullo et al., 2005 Chromosome evolution and improved cytogenetic maps of the Y chromosome in cattle, zebu, river buffalo, sheep and goat. Chromosome Research 13: 349-355.

Dilsaver, L. M., 2000 America's national park system: The critical documents. Rowman & Littlefield Publishers, Lanham, MD.
http://www.nps.gov/parkhistory/online_books/anps/anps_toc.htm

Douglas, K. C., N. D. Halbert, C. Kolenda, C. Childers, D. L. Hunter et al., 2011 Complete mitochondrial DNA sequence analysis of *Bison bison* and bison–cattle hybrids: function and phylogeny. Mitochondrion 11: 166-175.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, K. Billis et al., 2013 Ensembl 2014. Nucleic Acids Research 41: D48-55.

Flores, D., 1991 Bison ecology and bison diplomacy: the southern plains from 1800 to 1850. The Journal of American History 78: 465-485.

Freese, C. H., K. E. Aune, D. P. Boyd, J. N. Derr, S. C. Forrest et al., 2007 Second chance for the plains bison. Biological Conservation 136: 175-184.

Garretson, M. S., 1938 The American bison: the story of its extermination as a wild species and its restoration under federal protection. New York Zoological Society, New York, NY.

Geist, V., 1991 Phantom subspecies: the wood bison *Bison bison" athabascae"* Rhoads 1897 is not a valid taxon, but an ecotype. Arctic 44: 283-300.

Gnerre, S., E. S. Lander, K. Lindblad-Toh and D. B. Jaffe, 2009 Assisted assembly: how to improve a de novo genome assembly by using related species. Genome Biology 10: R88.

Halbert, N. D., 2003 The utilization of genetic markers to resolve modern management issues in historic bison populations: implications for species conservation. PhD Dissertation (Genetics). Texas A&M University, College Station, TX.

Halbert, N. D., T. Raudsepp, B. P. Chowdhary and J. N. Derr, 2004 Conservation genetic analysis of the Texas state bison herd. Journal of Mammalogy 85: 924-931.

Halbert, N. D., T. J. Ward, R. D. Schnabel, J. F. Taylor and J. N. Derr, 2005 Conservation genomics: disequilibrium mapping of domestic cattle chromosomal segments in North American bison populations. Molecular Ecology 14: 2343-2362.

Halbert, N. D., and J. N. Derr, 2008 Patterns of genetic variation in US federal bison herds. Molecular Ecology 17: 4963-4977.

Haley, J. E., 1949 Charles Goodnight: Cowman and Plainsman. University of Oklahoma Press, Norman, OK.

Hall, E. R.,1981 The mammals of North America. John Wiley and Sons, New York, NY.

Hornaday, W. T., 1887 The Extermination of the American Bison. National Museum, Washington D.C.

Hornaday, W. T., 1913 Our vanishing wild life: its extermination and preservation. New York Zoological Society, New York, NY.

Huang, D. W., B. T. Sherman and R. A. Lempicki, 2008 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4: 44-57.

Janecek, L. L., R. L. Honeycutt, R. M. Adkins and S. K. Davis, 1996 Mitochondrial gene sequences and the molecular systematics of the artiodactyl subfamily Bovinae. Molecular Phylogenetics and Evolution 6: 107-119.

Krumbiegel, I., and G. G. Sehm, 1989 The geographic variability of the Plains Bison. A reconstruction using the earliest European illustrations of both subspecies*. Archives of Natural History 16: 169-190.

Lee, T.-H., H. Guo, X. Wang, C. Kim and A. Paterson, 2014 SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15: 1-6.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079.

Li, H., 2011 Improving SNP discovery by base alignment quality. Bioinformatics 27: 1157-1158.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

McDonald, J. N., 1981 North American bison: their classification and evolution. University of California Press, Berkeley, CA.

McHugh, T., 1972 The time of the buffalo. University of Nebraska Press, Lincoln, NE.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis et al., 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303.

Meagher, M., 1986 *Bison bison*. Mammalian Species 266: 1-8.

Meagher, M. M., 1973 The bison of Yellowstone National Park. National Park Service, Washington, D.C.

Novakowski, N., and A. W. F. Banfield, 1960 The survival of the wood bison (*Bison bison athabascae* Rhoads) in the Northwest Territories. Natural History Papers 8. National Museum of Canada, Ottawa, Canada.

Park, S. D. E., 2001 Trypanotolerance in West African cattle and the population genetic effects of selection. PhD Dissertation. University of Dublin, Ireland.

Pertoldi, C., J. M. Wójcik, M. Tokarska, A. Kawałko, T. N. Kristensen et al., 2010 Genome variability in European and American bison detected using the BovineSNP50 BeadChip. Conservation Genetics 11: 627-634.

Polziehn, R. O., C. Strobeck, J. Sheraton and R. Beech, 1995 Bovine mtDNA discovered in North American bison populations. Conservation Biology: 1638-1643.

Polziehn, R. O., C. Strobeck, R. Beech and J. Sheraton, 1996 Genetic relationships among North American bison populations. Canadian Journal of Zoology 74: 738-749.

Potter, B., S. Gerlach, C. Gates, D. Boyd, G. Oetelaar et al., 2010 History of bison in North America. American Bison: Status Survey and Conservation Guidelines: 5-12.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Raudsepp, T., and B. P. Chowdhary, 2008 FISH for mapping single copy genes, pp. 31-49 in Phylogenomics. Springer.

Roe, F. G., 1970 The North American Buffalo: A critical study of the species in its wild state. University of Toronto Press, Toronto, Canada.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 Molecular cloning. Cold spring harbor laboratory press New York, NY.

Sanderson, E. W., K. H. Redford, B. Weber, K. Aune, D. Baldes et al., 2008 The ecological future of the North American Bison: conceiving long-term, large-scale conservation of wildlife. Conservation Biology 22: 252-266.

Schnabel, R. D., T. Ward and J. Derr, 2000 Validation of 15 microsatellites for parentage testing in North American bison, *Bison bison*, and domestic cattle. Animal Genetics 31: 360-366.

Seabright, M., 1972 Human chromosome banding. The Lancet 299: 967.

Seton, E. T., 1937 Lives of game animals. Literary Guild of America, New York, NY.

Simpson, G. G., 1961 Principles of animal taxonomy. Columbia University Press, New York, NY.

Soderlund, C., W. Nelson, A. Shoemaker and A. Paterson, 2006 SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Research 16: 1159-1168.

Swepston, D. A., 2001 Texas State Bison Herd, 2nd annual report, pp. 8. Texas Parks & Wildlife, Austin, TX.

Thibaud-Nissen F, A. Souvorov, T. Murphy, M. DiCuccio, and P. Kitts, 2013 Eukaryotic Genome Annotation Pipeline. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); http://www.ncbi.nlm.nih.gov/books/NBK169439/

van Camp, J., 1989 A surviving herd of endangered wood bison at Hook Lake, N.W.T.? Arctic 42: 314-322.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, et al., 2013 From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 11:11.10:11.10.1–11.10.33.

van Gelder, R. G., 1977 Mammalian hybrids and generic limits. American Museum Novitates 2635: 1-25.

Verkaar, E. L. C., H. Vervaecke, C. Roden, L. Romero Mendoza, M. W. Barwegen et al., 2003 Paternally inherited markers in bovine hybrid populations. Heredity 91: 565-569.

Ward, T. J., J. P. Bielawski, S. K. Davis, J. W. Templeton and J. N. Derr, 1999 Identification of domestic cattle hybrids in wild cattle and bison species: a general approach using mtDNA markers and the parametric bootstrap. Animal Conservation 2: 51-57.

Ward, T. J., 2000 An Evaluation of the Outcome of Interspecific Hybridization Events Coincident With a Dramatic Demographic Decline in North American Bison, pp. 116 in Ph.D. Dissertation (Genetics). Texas A&M University, College Station, TX.

Ward, T. J., L. C. Skow, D. S. Gallagher, R. D. Schnabel, C. A. Null et al., 2001 Differential introgression of uniparentally inherited markers in bison populations with hybrid ancestries. Animal Genetics 32: 89-91.

Zimin, A., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg et al., 2013 The MaSuRCA genome Assembler. Bioinformatics 29 (21): 2669-2677.

APPENDIX

The appendix tables contain the results of the DAVID analyses, SyMap synteny analysis, and variant effects found from SNPEff for each individual population. This should be used to reference more in depth results from the analyses of this dissertation work when cited.