이학박사 학위논문

Characterization and Population Differentiation of Structural Variation in *Bos taurus* and *Sus scrofa*

소와 돼지의 구조 변이의 특성 및 집단 간 차이 연구

2023 년 8월

서울대학교 대학원

협동과정 생물정보학전공

장 지 성

Characterization and Population Differentiation of Structural Variation in *Bos taurus* and *Sus scrofa*

지도 교수 김 희 발

서울대학교 대학원

협동과정 생물정보학전공

장지성

장지성의 이학박사 학위논문을 인준함

2023년 8월

위 원 장 ______ 정 충 원 ______ **(인)**

부위원장 ______ 김 희 발 ______ **(인)**

위 원 ______ 유 경 록 ______ **(인)**

위 원 ______ 조 서 애 ______ **(인)**

위 원 ______ 이 원 석 ______ **(인)**

Abstract

Characterization and Population Differentiation of Structural Variation in *Bos taurus* and *Sus scrofa*

Jisung Jang
Interdisciplinary Program in Bioinformatics
The Graduate School Seoul National University

Structural variation (SV) is a class of genomic alteration that involves segments of DNA longer than 1 kb. SVs can affect gene expression, function, and evolution, and are associated with various phenotypes and diseases. In this study, I investigated the population differentiation and characteristics of SVs in cattle and swine, two important domesticated animals with complex evolutionary histories. I used various genomic approaches to analyze SV in three research chapters that focused on copy number variation (CNV), a type of SV that involves deletion or duplication of DNA segments. Literature review about SV and approaches for identifying SV are summarized in the first chapter. The second chapter examined the population differentiated CNV of *Bos taurus*, *Bos indicus*, and their African hybrids, revealing the impact of hybridization and selection on CNV diversity. The third chapter compared the CNV between Eurasian wild boar and domesticated pig populations, uncovering the signatures of domestication and adaptation on CNV patterns. The fourth chapter presented the chromosome-level genome assembly of Hanwoo, a Korean native cattle breed, and the pangenome graph of 14 B. taurus assemblies. The study identified Hanwoo-specific regions and structural variants that may be related to phenotypic traits and adaptation. These chapters collectively demonstrated the power and utility of population genetics of SVs for studying the evolution and disease of cattle and swine and provided valuable resources and insights for future research.

Keyword: Population genetics, Structural variation, Copy number variation, Evolution

Student Number: 2016-28977

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1. Structural variation and population genetics

Structural variation (SV) is a type of genomic alteration that involves segments of DNA longer than 1 kb (Collins et al., 2020). SVs can be classified into unbalanced rearrangements, such as copy-number variants (CNVs), which result in gains or losses of DNA, and balanced rearrangements, such as inversions and translocations, which occur without corresponding dosage changes (Collins et al., 2020). SVs can affect protein-coding genes and cis-regulatory elements, and have profound consequences for genome evolution and function (Collins et al., 2020). SVs can also contribute to human diseases, such as autism, schizophrenia, and cancer (Collins et al., 2020).

Population genetics is the study of the distribution and dynamics of genetic variation within and between populations (Conrad & Hurles, 2007). Population genetics can help us understand the origins and impacts of SVs in a species by linking evolutionary themes. For example, population genetics can reveal how natural selection, genetic drift, recombination, migration, and population demography influence the frequency and diversity of SVs across different geographical regions and ethnic groups (Conrad & Hurles, 2007). Population genetics can also identify SVs that are outliers or signatures of adaptation or disease susceptibility (Conrad & Hurles, 2007).

1.2. Methods used to study structural variation of different populations in a same species.

To study the population genetics of SVs, various methods have been developed to detect and characterize SVs from genomic data. These methods can be broadly categorized into three types: array-based methods, sequence-based methods (Collins et al., 2020). Array-based methods use microarrays to measure the relative hybridization intensity of genomic DNA from different individuals or samples. These methods can detect CNVs with high resolution and accuracy, but they are limited by the availability and design of probes on the array (Collins et al., 2020). Sequence-based methods use next-generation sequencing (NGS) data to identify SVs by comparing the read depth, read pair, or split-read information of genomic DNA from different individuals or samples. These methods can detect a wide range of SV types and sizes with high sensitivity and specificity, but they require high sequencing coverage and computational resources (Collins et al., 2020). Hybrid methods combine array-based and sequence-based approaches to leverage the advantages of both technologies. These methods can provide comprehensive and reliable SV detection and genotyping across diverse populations (Collins et al., 2020).

In this thesis, I use sequence-based methods to study the population genetics of SVs in two different species: cattle (*Bos taurus*) and pigs (*Sus scrofa*). I focus on CNVs as a major class of SVs that affect gene dosage and expression. I compare the CNV profiles of different populations within each species to investigate the evolutionary forces shaping their genomic diversity. I also explore the functional implications of CNVs for phenotypic variation and disease resistance.

In the process of using sequence-based SVs, I encountered reference-biased problems. To overcome this, I performed reference genome assembly of Korean indigenous cattle, Hanwoo, and constructed a pangenome by collecting all existing high-quality assemblies of *bos taurus*. Moreover, I visualized the SVs that are expected to have evolutionary significance by representing the pangenome as a multi-assembly graph.

This chapter was published in *BMC genomics*

as a partial fulfillment of Jisung Jang's Ph.D program,

Chapter 2. Population differentiated copy number variation of Bos taurus, Bos indicus and their African hybrids

2.1. Abstract

Background

CNV comprises a large proportion in cattle genome and is associated with various traits. However, there were few population-scale comparison studies on cattle CNV.

Results

Here, autosome-wide CNVs were called by read depth of NGS alignment result and copy number variation regions (CNVRs) defined from 102 Eurasian taurine (EAT) of 14 breeds, 28 Asian indicine (ASI) of 6 breeds, 22 African taurine (AFT) of 2 breeds, and 184 African humped cattle (AFH) of 17 breeds. The copy number of every CNVRs were compared between populations and CNVRs with population differentiated copy numbers were sorted out using the pairwise statistics V_{ST} and *Kruskal-Wallis* test. Three hundred sixty-two of CNVRs were significantly differentiated in both statistics and 313 genes were located on the population differentiated CNVRs.

Conclusion

For some of these genes, the averages of copy numbers were also different between populations, and these may be candidate genes under selection. These include olfactory receptors, pathogen-resistance, parasite-resistance, heat tolerance and productivity related genes. Furthermore, breed- and individual-level comparison was performed using the presence or copy number of the autosomal CNVRs. my findings were based on identification of CNVs from short Illumina reads of 336 individuals and 39 breeds, which to my knowledge is the largest dataset for this type of analysis and revealed important CNVs that may play a role in cattle adaption to various environments.

2.2. Introduction

Cattle (Bos taurus) has been an invaluable animal providing livestock products such as milk, meat, leather and acting as a draft animal for cultivation and transportation since the domestication of extinct wild aurochs (Bos primigenius) (Magee et al., 2014). The two subspecies of Bos taurus taurus, taurine (B. t. taurus) and zebu (B. t. indicus) were brought about after bifurcation in 335,000 YBP, and were domesticated independently in different time and location (Achilli et al., 2009; Loftus et al., 1994). Archaeological and genomic evidence indicate that the taurine was domesticated approximately 10,000 YBP in Fertile Crescents and the zebu was domesticated 8,000 YBP in Indus Valley (Ajmone-Marsan et al., 2010; Chen et al., 2010; Vigne, 2011). The domesticated cattle populations were dispersed quickly after domestication along with the migration of pastoralists (Ajmone-Marsan et al., 2010). Their adaption to various local environments, artificial selection and introgression gave rise to genetically and phenotypically diversified modern cattle breeds (Decker et al., 2014).

Genome-wide variations such as SNPs and small INDELs of cattle were identified in previous studies (Consortium, 2009; Hayes et al., 2014). These small variations have been studied for understanding cattle evolution including population structure, selection, demographic history, and introgression (Decker et al., 2014; Kim et al., 2017; Kim et al., 2020). In case of structural variation, a large proportion in the genome is comprised of CNVs which have great effects on changing of gene structure, dosage, and expression level (Keel et al., 2016; Zhang et al., 2009). Despite its potentially high functional effects and abundance in the genome, insufficient data, and absence of standards in detection and downstream analysis make understanding of CNVs and their impact in cattle genome difficult. However, recent release of the high quality cattle genome assemblies such as ARS-UCD1.2, UOA_Angus_1 and UOA_Brahman_1 make NGS based CNV study available and more credible (Low et al., 2020; Rosen et al., 2020)…The CNV calling based on short read mapping became able to detect rare or novel variants, expanded target region to genome-wide, and improved resolution of the location (Mielczarek et al., 2018).

Here, I detected genome-wide CNVs of 336 individuals in 39 global cattle breeds including Eurasian taurine, Asian indicine and African cattle, and 2 individuals of African buffalo (Syncerus caffer caffer) using NGS read mapping. This is the largest number of breeds and individuals used in an NGS read mapping based cattle CNV study, including, notably, 19 breeds of African cattle that have not been well understood in terms of their CNVs. CNVs were defined from paired-end mapping result of short reads produced by Illumina HiSeq or NovaSeq platform. I performed population genetics survey on autosomal copy number variation regions (CNVRs). Hierarchical clustering of CNVRs from all individuals were compared to geographical origins and breeds. CNVRs with population differentiated copy number were identified by pairwise comparison of variance and rank based statistics. Population differentiated CNVRs overlapping genes were functionally annotated and suggested as candidate genes associated with selection and adaptation.

2.3. Materials and Methods

2.3.1. Sample collection

The study population consisted of 336 individuals of 39 cattle breeds and 2 individuals of African Buffalo (*Syncerus caffer*, AFB). Most of individuals except for 10 Bale, 10 Bagaria, 10 Semien and 5 Afar were included in previous SNP-based study by Kim et al. (Kim et al., 2020). Names of common individuals here followed the names used in the forementioned study (Kim et al., 2020). Breeds of the two subspecies *bos taurus taurus* and *bos taurus indicus* were collected from Europe, Asia and Africa. Humpless taurine and the crossbreeds such as Sanga (*Bos taurus taurus* x *Bos taurus indicus*) and Zenga cattle (Sanga x *Bos taurus indicus*) were collected from Africa. The 39 *bos taurus* breeds were classified into four groups by their original region and subspecies as following: i) 102 individuals of European and Asian taurine (EAT) which included 10 Angus, 10 Holstein, 18 Hereford, 10 Jersey, 11 Simmental, 5 Eastern Finn, 5 Western Finn, 3 Maremmana, 2 Sayaguesa, 2 Pajuna, 1 Limia, 1 Maronesa, 1 Podolica and 23 Hanwoo; ii) 28 individuals of Asian indicine (ASI) which included 16 Brahman, 6 Nelore, 3 Gir, 1 Hariana, 1 Sahiwal and 1 Tharparkar; iii) 22 individuals of African tarurine (AFT) which included 9 Muturu and 13 N'Dama; and iv) 184 individuals of African humped cattle (AFH) which included African zebu and the crossbreeds such as sanga (zebu x taurine) and zenga (zebu x sanga). The African zebu consisted of 10 Arsi, 10 Bagaria, 10 Bale, 9 Barka, 20 Butana, 10 EthiopianBoran, 10 Goffa, 13 Kenana, 10 KenyaBoran, 10 Mursi, 9 Ogaden and 10 Semien. Sanga consisted of 14 Afar, 10 Ankole and 9 Sheko, and Zenga consisted of 9 Fogera and 11 Horro. Genomes of all individuals were sequenced by Illumina paired-end library and their additional information is described on Table 2.1. The publicly available sequences were downloaded from SRA with following project accession numbers; PRJNA574857 (Afar, African Buffalo, Arsi, Barka, Butana, Ethiopian Boran, Fogera, Goffa, Horro, Kenana, Mursi, N'Dama, Sheko), PRJNA318087 (Angus, Ankole, Jersey, Kenya, Boran, Kenana, N'Dama, and Ogaden), PRJNA514237 (Limia, Maremmana, Maronesa, Pajuna, Podolica, and Sayaguesa), PRJNA324822 (Brahman), PRJNA343262 (Brahman, Gir, Hereford, Nelore, and Simmental), PRJNA432125 (Brahman), PRJEB28185 (Eastern Finn, and Western Finn), PRJNA210523 (Hanwoo), PRJNA379859 (Hariana, Sahiwal, and Thaparkar), PRJNA210521 (Holstein), PRJNA386202 (Muturu), and PRJNA507259 (Nelore)

2.3.2. Whole genome sequence alignment

After quality control checking of raw reads using FastQC-0.11.8 (Andrews, 2010), adapter and low-quality bases of reads were trimmed by Trimmomatic-0.39 (Bolger et al., 2014). After check result of trimming and quality of trimmed reads, the trimmed reads were mapped using BWA-0.7.17 MEM (Li & Durbin, 2009) to reference genome ARS-UCD1.2 with Btau5.0.1 Y chromosome assembly. The output of sequence alignment map (SAM) was sorted, indexed, and compressed to binary format (BAM) by Samtools-1.9 (Liu et al., 2009). The duplicates in BAM were marked using Picard 2.20.2 MarkDuplicates (https://broadinstitute.github.io/picard/) and the marked BAM files were used as input of variant calling. The alignment rate, coverage and mean depth were calculated using Sambamba (Tarasov et al., 2015).

Table 2.1. Sample information and alignment statistics.

Group	Name	Breed	Sex	Accession	MappingRate (%)	Coverage (%)	MeanDepth	Instrument
AFS	AFA01	Afar	F	SAMN15514550	99.83	95.19	10.44	HiSeq2000
AFS	AFA02	Afar	F	SAMN15514551	99.83	95.08	9.44	HiSeq2000
AFS	AFA03	Afar	F	SAMN15514552	99.78	95.12	9.83	HiSeq2000
AFS	AFA04	Afar	F	SAMN15514553	99.69	95.18	8.99	HiSeq2000
AFS	AFA05	Afar	F	SAMN15514554	99.81	95.00	9.91	HiSeq2000
AFS	AFA06	Afar	F	SAMN15514555	99.70	95.11	8.65	HiSeq2000
AFS	AFA07	Afar	F	SAMN15514556	99.81	95.15	9.46	HiSeq2000
AFS	AFA08	Afar	M	SAMN15514557	99.68	95.28	8.67	HiSeq2000
AFS	AFA09	Afar	M	SAMN15514558	99.80	95.30	9.51	HiSeq2000
AFS	AFA10	Afar	NA	SAMN17765866	99.45	95.25	14.95	HiSeq2500
AFS	AFA11	Afar	NA	SAMN17765867	99.42	95.25	13.61	HiSeq2500
AFS	AFA12	Afar	NA	SAMN17765868	99.42	95.41	12.26	HiSeq2500
AFS	AFA13	Afar	NA	SAMN17765869	99.51	95.24	13.27	HiSeq2500
AFS	AFA14	Afar	NA	SAMN17765870	99.50	95.35	15.16	HiSeq2500
AFB	AFB01	AfricanBuffalo	NA	SAMN15514475	98.56	92.33	18.49	HiSeq2000
AFB	AFB02	AfricanBuffalo	NA	SAMN15514476	99.37	92.21	19.38	HiSeq2000
EUT	ANG01	Angus	NA	SAMN04978232	99.85	95.66	8.89	HiSeq2000

EUT	ANG02	Angus	NA	SAMN04978233	99.89	95.41	9.56	HiSeq2000
EUT	ANG03	Angus	NA	SAMN04978234	99.82	95.67	9.54	HiSeq2000
EUT	ANG04	Angus	NA	SAMN04978235	99.89	95.63	9.77	HiSeq2000
EUT	ANG05	Angus	NA	SAMN04978238	99.91	95.37	9.86	HiSeq2000
EUT	ANG06	Angus	NA	SAMN04978239	99.84	95.76	10.47	HiSeq2000
EUT	ANG07	Angus	NA	SAMN04978240	99.86	95.46	10.36	HiSeq2000
EUT	ANG08	Angus	NA	SAMN04978241	99.88	95.31	10.40	HiSeq2000
EUT	ANG09	Angus	NA	SAMN04978236	99.85	95.75	8.00	HiSeq2000
EUT	ANG10	Angus	NA	SAMN04978237	99.87	95.27	6.74	HiSeq2000
AFS	ANK01	Ankole	NA	SAMN04545540	99.78	95.31	8.24	HiSeq2000
AFS	ANK02	Ankole	NA	SAMN04545541	99.83	95.26	7.57	HiSeq2000
AFS	ANK03	Ankole	NA	SAMN04545542	91.75	92.03	5.35	HiSeq2000
AFS	ANK04	Ankole	NA	SAMN04545543	99.81	95.32	8.47	HiSeq2000
AFS	ANK05	Ankole	NA	SAMN04545544	99.83	95.42	8.41	HiSeq2000
AFS	ANK06	Ankole	NA	SAMN04545545	99.85	95.25	8.63	HiSeq2000
AFS	ANK07	Ankole	NA	SAMN04545546	99.82	95.22	8.51	HiSeq2000
AFS	ANK08	Ankole	NA	SAMN04545547	99.86	95.27	8.16	HiSeq2000
AFS	ANK09	Ankole	NA	SAMN04545548	99.83	95.29	8.07	HiSeq2000
AFS	ANK10	Ankole	NA	SAMN04545549	99.81	95.30	8.05	HiSeq2000
AFI	ARS01	Arsi	F	SAMN15514477	99.79	95.04	9.12	HiSeq2000
AFI	ARS02	Arsi	F	SAMN15514478	99.83	95.10	9.62	HiSeq2000

AFI	ARS03	Arsi	F	SAMN15514479	99.83	95.19	9.98	HiSeq2000
AFI	ARS04	Arsi	F	SAMN15514480	99.66	95.21	9.20	HiSeq2000
AFI	ARS05	Arsi	F	SAMN15514481	99.87	95.15	9.47	HiSeq2000
AFI	ARS06	Arsi	F	SAMN15514482	99.64	95.14	8.80	HiSeq2000
AFI	ARS07	Arsi	M	SAMN15514483	99.80	95.23	10.24	HiSeq2000
AFI	ARS08	Arsi	M	SAMN15514484	99.83	95.30	9.15	HiSeq2000
AFI	ARS09	Arsi	M	SAMN15514485	99.59	95.30	8.37	HiSeq2000
AFI	ARS10	Arsi	M	SAMN15514486	99.84	95.35	10.00	HiSeq2000
AFI	BAG01	Bagaria	F	SAMN17765871	99.61	94.20	23.38	HiSeq2500
AFI	BAG02	Bagaria	F	SAMN17765872	99.63	94.19	23.15	HiSeq2500
AFI	BAG03	Bagaria	F	SAMN17765873	99.63	94.19	23.61	HiSeq2500
AFI	BAG04	Bagaria	F	SAMN17765874	99.66	94.16	23.51	HiSeq2500
AFI	BAG05	Bagaria	F	SAMN17765875	99.81	95.36	24.32	HiSeq2500
AFI	BAG06	Bagaria	F	SAMN17765876	99.66	94.11	21.61	HiSeq2500
AFI	BAG07	Bagaria	F	SAMN17765877	99.63	94.13	21.41	HiSeq2500
AFI	BAG08	Bagaria	M	SAMN17765878	99.63	94.24	21.03	HiSeq2500
AFI	BAG09	Bagaria	F	SAMN17765879	99.74	95.40	26.99	HiSeq2500
AFI	BAG10	Bagaria	F	SAMN17765880	99.63	94.24	25.42	HiSeq2500
AFI	BAL01	Bale	F	SAMN17765881	99.60	94.13	21.24	HiSeq2500
AFI	BAL02	Bale	F	SAMN17765882	99.59	94.15	21.09	HiSeq2500
AFI	BAL03	Bale	F	SAMN17765883	99.64	94.18	22.74	HiSeq2500

AFI	BAL04	Bale	F	SAMN17765884	99.73	94.15	21.94	HiSeq2500
AFI	BAL05	Bale	F	SAMN17765885	99.66	94.17	23.52	HiSeq2500
AFI	BAL06	Bale	M	SAMN17765886	99.61	94.25	21.21	HiSeq2500
AFI	BAL07	Bale	F	SAMN17765887	99.73	94.23	25.63	HiSeq2500
AFI	BAL08	Bale	F	SAMN17765888	99.70	94.21	26.62	HiSeq2500
AFI	BAL09	Bale	F	SAMN17765889	99.66	94.25	26.13	HiSeq2500
AFI	BAL10	Bale	F	SAMN17765890	99.71	94.14	23.64	HiSeq2500
AFI	BAR01	Barka	NA	SAMN15514487	99.67	95.06	8.30	HiSeq2000
AFI	BAR02	Barka	NA	SAMN15514488	99.77	95.15	9.65	HiSeq2000
AFI	BAR03	Barka	NA	SAMN15514489	99.81	95.08	9.53	HiSeq2000
AFI	BAR04	Barka	NA	SAMN15514490	99.82	95.06	9.32	HiSeq2000
AFI	BAR05	Barka	NA	SAMN15514491	99.73	95.22	9.06	HiSeq2000
AFI	BAR06	Barka	NA	SAMN15514492	99.58	95.17	10.12	HiSeq2000
AFI	BAR07	Barka	NA	SAMN15514493	99.54	95.18	8.92	HiSeq2000
AFI	BAR08	Barka	NA	SAMN15514494	99.80	95.04	9.69	HiSeq2000
AFI	BAR09	Barka	NA	SAMN15514495	99.61	95.21	10.75	HiSeq2000
ASI	BRA04	Brahman	M	SAMN05788495	99.73	93.56	6.27	HiSeq2000, GAIIx
ASI	BRA06	Brahman	M	SAMN08435316	99.02	94.73	10.60	HiSeqXTen
ASI	BRA07	Brahman	M	SAMN08435281	99.07	95.12	14.65	HiSeqXTen
ASI	BRA08	Brahman	M	SAMN08435282	99.07	95.15	14.16	HiSeqXTen
ASI	BRA09	Brahman	M	SAMN08435279	99.10	95.16	14.67	HiSeqXTen

ASI	BRA10	Brahman	M	SAMN08435280	99.15	95.15	14.77	HiSeqXTen
ASI	BRA11	Brahman	M	SAMN08435317	99.16	94.83	10.90	HiSeqXTen
ASI	BRA12	Brahman	M	SAMN08435327	99.58	95.04	11.72	HiSeqXTen
ASI	BRA13	Brahman	M	SAMN08435322	99.37	95.07	10.26	HiSeqXTen
ASI	BRA14	Brahman	M	SAMN08435324	99.08	95.18	16.08	HiSeqXTen
ASI	BRA15	Brahman	M	SAMN08435323	99.65	94.64	11.07	HiSeqXTen
ASI	BRA16	Brahman	M	SAMN05216066	99.77	95.39	9.98	NextSeq550
ASI	BRA17	Brahman	M	SAMN05216067	99.77	95.57	11.84	NextSeq550
ASI	BRA18	Brahman	M	SAMN05216068	99.72	95.46	10.44	NextSeq550
ASI	BRA19	Brahman	M	SAMN05216069	99.73	95.49	11.82	NextSeq550
ASI	BRA20	Brahman	M	SAMN05216070	99.77	95.59	13.15	NextSeq550
AFI	BUT01	Butana	NA	SAMN15514496	99.73	95.10	9.99	HiSeq2000
AFI	BUT02	Butana	NA	SAMN15514497	99.65	95.16	8.78	HiSeq2000
AFI	BUT03	Butana	NA	SAMN15514498	99.59	95.16	8.87	HiSeq2000
AFI	BUT04	Butana	NA	SAMN15514499	99.55	95.25	8.99	HiSeq2000
AFI	BUT05	Butana	NA	SAMN15514500	99.66	95.18	9.02	HiSeq2000
AFI	BUT06	Butana	NA	SAMN15514501	99.69	95.20	8.77	HiSeq2000
AFI	BUT07	Butana	NA	SAMN15514500	99.72	95.32	10.19	HiSeq2000
AFI	BUT08	Butana	NA	SAMN15514500	99.62	95.10	9.53	HiSeq2000
AFI	BUT09	Butana	NA	SAMN15514500	99.67	95.17	8.95	HiSeq2000
AFI	BUT10	Butana	NA	SAMN15514500	99.66	95.16	9.20	HiSeq2000

AFI	BUT11	Butana	NA	SAMN15514500	99.70	95.20	9.33	HiSeq2000
AFI	BUT12	Butana	NA	SAMN15514500	99.70	95.11	9.80	HiSeq2000
AFI	BUT13	Butana	NA	SAMN15514500	99.65	95.15	8.76	HiSeq2000
AFI	BUT14	Butana	NA	SAMN15514500	99.77	95.07	10.01	HiSeq2000
AFI	BUT15	Butana	NA	SAMN15514510	99.77	94.82	10.09	HiSeq2000
AFI	BUT16	Butana	NA	SAMN15514511	99.73	95.12	10.18	HiSeq2000
AFI	BUT17	Butana	NA	SAMN15514512	99.73	95.16	10.07	HiSeq2000
AFI	BUT18	Butana	NA	SAMN15514513	99.67	95.25	8.85	HiSeq2000
AFI	BUT19	Butana	NA	SAMN15514514	99.61	95.11	8.56	HiSeq2000
AFI	BUT20	Butana	NA	SAMN15514515	99.64	95.16	8.67	HiSeq2000
EUT	EAF01	Eastern Finn	F	SAMEA4827182	99.84	95.12	9.58	HiSeq2000
EUT	EAF02	Eastern Finn	F	SAMEA4827183	99.86	95.02	9.47	HiSeq2000
EUT	EAF03	Eastern Finn	F	SAMEA4827184	99.85	94.83	9.19	HiSeq2000
EUT	EAF04	Eastern Finn	F	SAMEA4827185	99.85	95.08	9.66	HiSeq2000
EUT	EAF05	Eastern Finn	F	SAMEA4827186	99.83	95.03	9.44	HiSeq2000
AFI	ETB01	Ethiopian Boran	F	SAMN15514516	99.82	95.12	10.21	HiSeq2000
AFI	ETB02	Ethiopian Boran	F	SAMN15514517	99.82	95.06	9.66	HiSeq2000
AFI	ETB03	Ethiopian Boran	F	SAMN15514518	99.84	95.15	9.39	HiSeq2000
AFI	ETB04	Ethiopian Boran	F	SAMN15514519	99.77	95.12	9.95	HiSeq2000
AFI	ETB05	Ethiopian Boran	F	SAMN15514520	99.75	95.15	9.20	HiSeq2000

AFI	ETB06	Ethiopian Boran	F	SAMN15514521	99.58	95.09	9.04	HiSeq2000
AFI	ETB07	Ethiopian Boran	F	SAMN15514522	99.71	95.08	8.41	HiSeq2000
AFI	ETB08	Ethiopian Boran	M	SAMN15514523	99.60	95.28	8.66	HiSeq2000
AFI	ETB09	Ethiopian Boran	M	SAMN15514524	99.65	95.31	8.84	HiSeq2000
AFI	ETB10	Ethiopian Boran	M	SAMN15514525	99.72	95.39	9.24	HiSeq2000
AFZ	FOG01	Fogera	F	SAMN15514571	99.81	95.15	9.78	HiSeq2000
AFZ	FOG02	Fogera	F	SAMN15514572	99.83	95.19	9.40	HiSeq2000
AFZ	FOG03	Fogera	F	SAMN15514573	99.62	95.01	8.80	HiSeq2000
AFZ	FOG04	Fogera	F	SAMN15514574	99.83	95.22	9.46	HiSeq2000
AFZ	FOG05	Fogera	F	SAMN15514575	99.83	95.20	10.43	HiSeq2000
AFZ	FOG06	Fogera	F	SAMN15514576	99.85	95.21	11.04	HiSeq2000
AFZ	FOG07	Fogera	F	SAMN15514577	99.83	95.16	9.64	HiSeq2000
AFZ	FOG08	Fogera	F	SAMN15514578	99.57	95.12	8.77	HiSeq2000
AFZ	FOG09	Fogera	M	SAMN15514579	99.81	95.38	9.89	HiSeq2000
ASI	GIR01	Gir	F	SAMN05788512	99.77	95.06	6.70	HiSeq2000
ASI	GIR02	Gir	M	SAMN05788513	99.81	95.22	9.79	HiSeq2000
ASI	GIR03	Gir	F	SAMN05788514	99.64	95.31	8.87	HiSeq2000
AFI	GOF01	Goffa	F	SAMN15514526	99.84	95.33	9.95	HiSeq2000
AFI	GOF02	Goffa	F	SAMN15514527	99.77	95.30	9.97	HiSeq2000
AFI	GOF03	Goffa	F	SAMN15514528	99.73	95.05	9.83	HiSeq2000

AFI	GOF04	Goffa	F	SAMN15514529	99.70	95.23	7.81	HiSeq2000
AFI	GOF05	Goffa	F	SAMN15514530	99.85	95.05	9.34	HiSeq2000
AFI	GOF06	Goffa	M	SAMN15514531	99.86	94.85	9.19	HiSeq2000
AFI	GOF07	Goffa	M	SAMN15514532	99.86	95.02	10.07	HiSeq2000
AFI	GOF08	Goffa	M	SAMN15514533	99.85	95.09	9.44	HiSeq2000
AFI	GOF09	Goffa	M	SAMN15514534	99.82	95.04	7.14	HiSeq2000
AFI	GOF10	Goffa	M	SAMN15514535	99.77	95.05	8.16	HiSeq2000
AST	HAN01	Hanwoo	NA	SAMN02225725	99.83	95.04	9.46	HiSeq2000
AST	HAN02	Hanwoo	NA	SAMN02225726	99.82	95.18	6.83	HiSeq2000
AST	HAN03	Hanwoo	NA	SAMN02225727	99.54	95.10	11.34	HiSeq2000
AST	HAN04	Hanwoo	NA	SAMN02225728	99.81	95.13	8.60	HiSeq2000
AST	HAN05	Hanwoo	NA	SAMN02225729	99.83	95.03	10.43	HiSeq2000
AST	HAN06	Hanwoo	NA	SAMN02225730	99.85	95.34	8.48	HiSeq2000
AST	HAN07	Hanwoo	NA	SAMN02225731	99.84	95.52	9.86	HiSeq2000
AST	HAN08	Hanwoo	NA	SAMN02225732	99.82	95.36	11.49	HiSeq2000
AST	HAN09	Hanwoo	NA	SAMN02225733	99.82	95.41	8.98	HiSeq2000
AST	HAN10	Hanwoo	NA	SAMN02225723	99.83	95.43	8.77	HiSeq2000
AST	HAN11	Hanwoo	NA	SAMN02225724	99.83	95.46	7.76	HiSeq2000
AST	HAN12	Hanwoo	NA	SAMN02225744	99.84	95.53	11.23	HiSeq2000
AST	HAN13	Hanwoo	NA	SAMN02225745	99.13	95.40	11.50	HiSeq2000
AST	HAN14	Hanwoo	NA	SAMN02225746	99.86	95.40	11.08	HiSeq2000

AST	HAN15	Hanwoo	NA	SAMN02225747	99.81	95.36	11.36	HiSeq2000
AST	HAN16	Hanwoo	NA	SAMN02225748	99.87	95.31	10.73	HiSeq2000
AST	HAN17	Hanwoo	NA	SAMN02225749	99.83	95.42	11.56	HiSeq2000
AST	HAN18	Hanwoo	NA	SAMN02225750	99.73	95.57	11.36	HiSeq2000
AST	HAN19	Hanwoo	NA	SAMN02225751	95.96	95.96	10.56	HiSeq2000
AST	HAN20	Hanwoo	NA	SAMN02225752	99.85	96.06	10.65	HiSeq2000
AST	HAN21	Hanwoo	NA	SAMN02225753	99.59	95.99	9.74	HiSeq2000
AST	HAN22	Hanwoo	NA	SAMN02225754	99.63	96.00	9.75	HiSeq2000
AST	HAN23	Hanwoo	NA	SAMN02225755	98.39	95.68	9.90	HiSeq2000
ASI	HAR03	Hariana	F	SAMN08862747	97.88	95.94	30.28	HiSeqXTen
EUT	HER01	Hereford	M	SAMN05788507	98.13	95.89	22.85	HiSeq2000, HiSeq2500
EUT	HER02	Hereford	M	SAMN05788531	99.42	95.99	19.57	HiSeq2500
EUT	HER03	Hereford	M	SAMN05788534	99.86	96.02	16.41	HiSeq2000
EUT	HER04	Hereford	M	SAMN05788535	97.84	95.88	19.04	HiSeq2000
EUT	HER05	Hereford	M	SAMN05788536	99.84	95.91	12.13	HiSeq2500
EUT	HER06	Hereford	M	SAMN05788537	99.84	95.99	17.01	HiSeq2500
EUT	HER07	Hereford	M	SAMN05788538	99.84	95.91	16.95	HiSeq2000
EUT	HER08	Hereford	M	SAMN05788539	99.88	95.88	18.81	HiSeq2000
EUT	HER09	Hereford	M	SAMN05788540	95.68	96.20	19.19	HiSeq2000
EUT	HER10	Hereford	M	SAMN05788555	99.26	95.81	17.24	HiSeq2000
EUT	HER11	Hereford	M	SAMN05788556	99.33	95.89	15.81	HiSeq2000

EUT	HER12	Hereford	M	SAMN05788557	99.83	95.95	16.38	HiSeq2500
EUT	HER13	Hereford	M	SAMN05788558	99.87	94.89	17.76	HiSeq2500
EUT	HER14	Hereford	M	SAMN05788559	99.85	94.85	14.53	HiSeq2000, HiSeq2500
EUT	HER15	Hereford	M	SAMN10940540	99.82	94.75	17.73	HiSeqXTen
EUT	HER16	Hereford	M	SAMN10940541	99.84	94.87	13.60	HiSeq2000, HiSeq2500
EUT	HER17	Hereford	M	SAMN10940542	99.82	95.11	14.58	HiSeq2000, HiSeq2500
EUT	HER18	Hereford	M	SAMN10940543	99.85	95.09	14.12	HiSeq2000, HiSeq2500
EUT	HOL01	Holstein	NA	SAMN02225734	99.80	95.02	8.45	HiSeq2000
EUT	HOL02	Holstein	NA	SAMN02225735	99.82	95.22	8.28	HiSeq2000
EUT	HOL03	Holstein	NA	SAMN02225736	96.42	94.93	9.50	HiSeq2000
EUT	HOL04	Holstein	NA	SAMN02225737	99.83	95.09	7.62	HiSeq2000
EUT	HOL05	Holstein	NA	SAMN02225738	99.76	94.95	10.00	HiSeq2000
EUT	HOL06	Holstein	NA	SAMN02225739	99.72	95.18	10.77	HiSeq2000
EUT	HOL07	Holstein	NA	SAMN02225740	99.81	95.12	10.78	HiSeq2000
EUT	HOL08	Holstein	NA	SAMN02225741	99.73	95.04	9.64	HiSeq2000
EUT	HOL09	Holstein	NA	SAMN02225742	98.60	95.10	9.96	HiSeq2000
EUT	HOL10	Holstein	NA	SAMN02225743	99.75	95.04	10.99	HiSeq2000
AFZ	HOR01	Horro	F	SAMN15514580	99.79	95.24	7.12	HiSeq2000
AFZ	HOR02	Horro	F	SAMN15514581	99.73	95.15	8.68	HiSeq2000
AFZ	HOR03	Horro	F	SAMN15514582	99.68	95.14	9.81	HiSeq2000
AFZ	HOR04	Horro	F	SAMN15514583	99.73	95.34	7.67	HiSeq2000

AFZ	HOR05	Horro	F	SAMN15514584	99.34	95.31	8.28	HiSeq2000
AFZ	HOR06	Horro	F	SAMN15514585	99.44	95.30	7.53	HiSeq2000
AFZ	HOR07	Horro	F	SAMN15514586	99.67	95.10	9.09	HiSeq2000
AFZ	HOR08	Horro	F	SAMN15514587	99.73	95.07	8.88	HiSeq2000
AFZ	HOR09	Horro	M	SAMN15514588	99.80	95.14	7.07	HiSeq2000
AFZ	HOR10	Horro	M	SAMN15514589	99.85	95.17	9.16	HiSeq2000
AFZ	HOR11	Horro	M	SAMN15514590	99.84	95.19	9.16	HiSeq2000
EUT	JER01	Jersey	NA	SAMN04978250	99.86	95.26	12.92	HiSeq2000
EUT	JER02	Jersey	NA	SAMN04978251	99.81	95.16	13.57	HiSeq2000
EUT	JER03	Jersey	NA	SAMN04978252	99.84	95.15	10.98	HiSeq2000
EUT	JER04	Jersey	NA	SAMN04978253	99.79	95.29	10.99	HiSeq2000
EUT	JER05	Jersey	NA	SAMN04978254	99.85	95.16	10.42	HiSeq2000
EUT	JER06	Jersey	NA	SAMN04978255	97.98	95.50	12.15	HiSeq2000
EUT	JER07	Jersey	NA	SAMN04978256	99.86	95.21	12.66	HiSeq2000
EUT	JER08	Jersey	NA	SAMN04978257	95.13	95.50	13.81	HiSeq2000
EUT	JER09	Jersey	NA	SAMN04978258	99.29	95.24	11.92	HiSeq2000
EUT	JER10	Jersey	NA	SAMN04978259	98.56	95.21	13.61	HiSeq2000
AFI	KEN01	Kenana	NA	SAMN15514536	99.75	91.75	8.53	HiSeq2000
AFI	KEN02	Kenana	NA	SAMN15514537	99.83	90.04	9.13	HiSeq2000
AFI	KEN03	Kenana	NA	SAMN15514538	99.84	95.18	9.56	HiSeq2000
AFI	KEN04	Kenana	NA	SAMN15514539	99.66	95.12	9.70	HiSeq2000

AFI	KEN05	Kenana	F	SAMN04545556	99.76	95.09	8.22	HiSeq2000
AFI	KEN06	Kenana	F	SAMN04545558	99.44	95.04	8.37	HiSeq2000
AFI	KEN07	Kenana	F	SAMN04545550	99.73	95.17	8.35	HiSeq2000
AFI	KEN08	Kenana	F	SAMN04545551	99.63	95.17	8.04	HiSeq2000
AFI	KEN09	Kenana	F	SAMN04545552	99.75	95.17	8.23	HiSeq2000
AFI	KEN10	Kenana	F	SAMN04545553	99.76	94.92	8.10	HiSeq2000
AFI	KEN11	Kenana	F	SAMN04545559	99.73	95.28	8.16	HiSeq2000
AFI	KEN12	Kenana	M	SAMN04545555	99.65	95.35	8.20	HiSeq2000
AFI	KEN13	Kenana	M	SAMN04545557	99.57	94.11	8.33	HiSeq2000
AFI	KEB01	KenyaBoran	NA	SAMN04545530	99.51	94.57	7.89	HiSeq2000
AFI	KEB02	KenyaBoran	NA	SAMN04545531	98.96	94.34	7.77	HiSeq2000
AFI	KEB03	KenyaBoran	NA	SAMN04545532	98.63	94.09	7.59	HiSeq2000
AFI	KEB04	KenyaBoran	NA	SAMN05862018	99.22	94.23	8.45	HiSeq2000
AFI	KEB05	KenyaBoran	NA	SAMN04545538	98.70	92.54	8.13	HiSeq2000
AFI	KEB06	KenyaBoran	NA	SAMN04545533	99.21	92.83	7.58	HiSeq2000
AFI	KEB07	KenyaBoran	NA	SAMN04545539	98.59	93.48	8.08	HiSeq2000
AFI	KEB08	KenyaBoran	NA	SAMN04545534	98.97	93.05	7.84	HiSeq2000
AFI	KEB09	KenyaBoran	NA	SAMN04545535	99.86	95.11	8.34	HiSeq2000
AFI	KEB10	KenyaBoran	NA	SAMN04545537	99.75	95.16	8.48	HiSeq2000
EUT	LIM01	Limia	M	SAMN10721581	99.72	95.21	8.05	HiSeq2000
EUT	MAM01	Maremmana	F	SAMN10721583	99.85	95.29	8.66	HiSeq2000

EUT	MAM02	Maremmana	F	SAMN10721583	99.84	95.27	8.47	HiSeq2000
EUT	MAM03	Maremmana	F	SAMN10721583	99.03	95.28	7.58	HiSeq2000
EUT	MAN01	Maronesa	M	SAMN10721580	99.79	95.28	7.53	HiSeq2000
AFI	MUR01	Mursi	F	SAMN15514540	99.78	95.33	9.39	HiSeq2000
AFI	MUR02	Mursi	F	SAMN15514541	99.79	95.33	8.63	HiSeq2000
AFI	MUR03	Mursi	F	SAMN15514542	99.85	95.30	9.50	HiSeq2000
AFI	MUR04	Mursi	F	SAMN15514543	99.69	95.25	10.59	HiSeq2000
AFI	MUR05	Mursi	F	SAMN15514544	99.81	95.27	8.40	HiSeq2000
AFI	MUR06	Mursi	F	SAMN15514545	99.78	95.27	8.26	HiSeq2000
AFI	MUR07	Mursi	F	SAMN15514546	99.79	95.17	8.21	HiSeq2000
AFI	MUR08	Mursi	F	SAMN15514547	99.80	95.23	6.82	HiSeq2000
AFI	MUR09	Mursi	M	SAMN15514548	99.80	95.19	9.16	HiSeq2000
AFI	MUR10	Mursi	M	SAMN15514549	99.81	95.19	9.12	HiSeq2000
AFT	MUT01	Muturu	NA	SAMN07135491	99.82	95.24	6.75	HiSeq2500
AFT	MUT02	Muturu	NA	SAMN07135492	99.86	95.15	7.79	HiSeq2500
AFT	MUT03	Muturu	NA	SAMN07135493	99.69	95.15	8.20	HiSeq2500
AFT	MUT04	Muturu	NA	SAMN07135494	99.75	95.28	7.58	HiSeq2500
AFT	MUT05	Muturu	NA	SAMN07135495	99.72	95.30	6.95	HiSeq2500
AFT	MUT06	Muturu	NA	SAMN07135496	99.85	95.58	5.15	HiSeq2500
AFT	MUT08	Muturu	NA	SAMN07135498	99.42	95.52	5.30	HiSeq2500
AFT	MUT09	Muturu	NA	SAMN07135499	96.21	95.27	6.59	HiSeq2500

AFT	MUT10	Muturu	NA	SAMN07135500	99.71	95.36	5.17	HiSeq2500
AFT	NDA01	N'Dama	NA	SAMN15514559	99.62	94.37	9.85	HiSeq2000
AFT	NDA02	N'Dama	NA	SAMN15514560	99.68	94.32	9.21	HiSeq2000
AFT	NDA03	N'Dama	NA	SAMN15514561	99.72	94.35	9.20	HiSeq2000
AFT	NDA04	N'Dama	NA	SAMN04545560	99.66	94.34	8.48	HiSeq2000
AFT	NDA05	N'Dama	NA	SAMN04545561	99.72	94.23	7.94	HiSeq2000
AFT	NDA06	N'Dama	NA	SAMN04545562	99.70	94.34	7.21	HiSeq2000
AFT	NDA07	N'Dama	NA	SAMN04545563	99.83	95.16	8.27	HiSeq2000
AFT	NDA08	N'Dama	NA	SAMN04545564	99.71	95.29	8.43	HiSeq2000
AFT	NDA09	N'Dama	NA	SAMN04545565	99.84	95.18	8.58	HiSeq2000
AFT	NDA10	N'Dama	NA	SAMN04545566	99.73	95.20	8.27	HiSeq2000
AFT	NDA11	N'Dama	NA	SAMN04545567	99.63	95.13	8.15	HiSeq2000
AFT	NDA12	N'Dama	NA	SAMN04545568	99.73	95.25	8.48	HiSeq2000
AFT	NDA13	N'Dama	NA	SAMN04545569	99.70	95.27	8.24	HiSeq2000
ASI	NEL01	Nelore	NA	SAMN05788520	99.84	95.19	5.63	HiSeq2000
ASI	NEL03	Nelore	NA	SAMN05788522	99.83	95.27	5.41	HiSeq2000
ASI	NEL05	Nelore	NA	SAMN05788524	99.78	95.75	6.61	HiSeq2000
ASI	NEL07	Nelore	NA	SAMN10486400	99.84	95.81	10.01	HiSeq2000
ASI	NEL08	Nelore	NA	SAMN10486401	99.85	95.78	7.16	HiSeq2000
ASI	NEL09	Nelore	NA	SAMN10486398	96.07	95.71	8.29	HiSeq2000
AFI	OGA01	Ogaden	NA	SAMN04545574	96.49	95.86	7.65	HiSeq2000

AFI	OGA02	Ogaden	NA	SAMN04545575	97.72	95.73	8.42	HiSeq2000
AFI	OGA03	Ogaden	NA	SAMN04545576	96.43	95.94	7.98	HiSeq2000
AFI	OGA04	Ogaden	NA	SAMN04545570	95.97	95.93	7.88	HiSeq2000
AFI	OGA05	Ogaden	NA	SAMN04545571	99.58	95.76	8.17	HiSeq2000
AFI	OGA06	Ogaden	NA	SAMN04545577	98.01	95.89	7.91	HiSeq2000
AFI	OGA07	Ogaden	NA	SAMN04545573	99.87	95.87	7.84	HiSeq2000
AFI	OGA08	Ogaden	NA	SAMN04545578	99.59	95.32	7.29	HiSeq2000
AFI	OGA09	Ogaden	NA	SAMN04545579	99.78	94.66	8.19	HiSeq2000
EUT	PAJ01	Pajuna	M	SAMN10721584	99.89	94.71	8.61	HiSeq2000
EUT	PAJ02	Pajuna	M	SAMN10721584	99.80	94.82	8.44	HiSeq2000
EUT	POD01	Podolica	F	SAMN10721582	99.18	94.73	7.77	HiSeq2000
ASI	SAH02	Sahiwal	F	SAMN08862748	99.81	94.64	18.19	HiSeqXTen
EUT	SAY01	Sayaguesa	F	SAMN10721579	91.70	95.29	7.81	HiSeq2000
EUT	SAY02	Sayaguesa	F	SAMN10721579	99.49	95.29	8.59	HiSeq2000
AFI	SEM01	Semien	F	SAMN17765891	99.62	94.19	25.35	HiSeq2500
AFI	SEM02	Semien	F	SAMN17765892	99.67	94.23	25.19	HiSeq2500
AFI	SEM03	Semien	M	SAMN17765893	99.67	94.32	23.40	HiSeq2500
AFI	SEM04	Semien	F	SAMN17765894	99.61	94.21	25.03	HiSeq2500
AFI	SEM05	Semien	M	SAMN17765895	99.62	94.37	25.22	HiSeq2500
AFI	SEM06	Semien	M	SAMN17765896	99.68	94.32	24.67	HiSeq2500
AFI	SEM07	Semien	M	SAMN17765897	99.72	94.35	26.43	HiSeq2500

AFI	SEM08	Semien	M	SAMN17765898	99.66	94.34	23.95	HiSeq2500
AFI	SEM09	Semien	F	SAMN17765899	99.72	94.23	25.41	HiSeq2500
AFI	SEM10	Semien	M	SAMN17765900	99.70	94.34	22.50	HiSeq2500
AFS	SHE01	Sheko	F	SAMN15514562	99.83	95.16	9.50	HiSeq2000
AFS	SHE02	Sheko	F	SAMN15514563	99.71	95.29	10.56	HiSeq2000
AFS	SHE03	Sheko	F	SAMN15514564	99.84	95.18	10.33	HiSeq2000
AFS	SHE04	Sheko	F	SAMN15514565	99.73	95.20	9.17	HiSeq2000
AFS	SHE05	Sheko	F	SAMN15514566	99.63	95.13	9.58	HiSeq2000
AFS	SHE06	Sheko	F	SAMN15514567	99.73	95.25	9.59	HiSeq2000
AFS	SHE07	Sheko	F	SAMN15514568	99.70	95.27	9.46	HiSeq2000
AFS	SHE08	Sheko	F	SAMN15514569	99.84	95.19	9.24	HiSeq2000
AFS	SHE09	Sheko	F	SAMN15514570	99.83	95.27	9.83	HiSeq2000
EUT	SIM01	Simmental	M	SAMN05788541	99.78	95.75	18.44	HiSeq2000, HiSeq2500
EUT	SIM02	Simmental	M	SAMN05788542	99.84	95.81	20.01	HiSeq2000, HiSeq2500
EUT	SIM03	Simmental	M	SAMN05788543	99.85	95.78	18.57	HiSeq2000, HiSeq2500
EUT	SIM04	Simmental	M	SAMN05788544	96.07	95.71	18.07	HiSeq2000
EUT	SIM05	Simmental	M	SAMN05788545	96.49	95.86	23.43	HiSeq2000, HiSeq2500
EUT	SIM06	Simmental	M	SAMN05788546	97.72	95.73	18.41	HiSeq2000
EUT	SIM07	Simmental	M	SAMN10940558	96.43	95.94	15.82	HiSeq2500
EUT	SIM08	Simmental	M	SAMN10940559	95.97	95.93	16.36	HiSeq2000, HiSeq2500
EUT	SIM09	Simmental	M	SAMN10940560	99.58	95.76	17.81	HiSeq2000, HiSeq2500

EUT	SIM10	Simmental	M	SAMN10940561	98.01	95.89	17.65	HiSeq2000, HiSeq2500
EUT	SIM11	Simmental	M	SAMN10940562	99.87	95.87	16.51	HiSeq2000
ASI	THA03	Tharparkar	F	SAMN08862749	99.59	95.32	14.29	HiSeqXTen
EUT	WEF01	WesternFinn	F	SAMEA4827187	99.78	94.66	7.31	HiSeq2000
EUT	WEF02	WesternFinn	F	SAMEA4827188	99.89	94.71	9.38	HiSeq2000
EUT	WEF03	WesternFinn	F	SAMEA4827189	99.80	94.82	9.46	HiSeq2000
EUT	WEF04	WesternFinn	F	SAMEA4827190	99.18	94.73	8.82	HiSeq2000
EUT	WEF05	WesternFinn	F	SAMEA4827191	99.81	94.64	8.96	HiSeq2000

2.3.3. CNV calling and CNVR definition

CNVs of all samples were called with a bin size of 200bp by CNVnator (Abyzov et al., 2011) and filtered with size (>1kb), p-value calculated using t-test statistics (<0.001) and fraction of reads with zero mapping quality (MQ0<0.5). The CNVs in unplaced scaffolds were removed. A 50% reciprocal overlap between filtered CNVs was defined as copy number variation region (CNVR) using 'CNV_overlap.py' script on GitHub (https://github.com/bjtrost/TCAG-WGS-CNV-workflow) (Trost et al., 2018). CNVRs found in more than two individuals were used for downstream analysis to minimize false-positive. (Pierce et al., 2018) Copy number of each CNVR was calculated based on aligned read depth and normalized using CNVnator. The normalized copy number of neutral region from diploid autosome was assumed to be 2.0.

2.3.4. Hierarchical clustering based on CNVR

To cluster individuals according to their CNV similarities, I made a vector of "0"s and "1"s for each individuals based on absence or presence of a specific CNVR in that particular individual. Hierarchical clustering with 1000 times of bootstrap resampling was performed on these vectors of every autosomal CNVR using pvclust with default option in R (Suzuki & Shimodaira, 2006). The 'correlation' and 'average' were used as distance measure and the agglomerative method, respectively. The approximately unbiased (AU) p-value was calculated by multiscale bootstrap resampling. The bootstrap probability (BP) p-value was calculated by ordinary bootstrap resampling based on unweighted pair-group average method (UPGMA).

2.3.5. Population differentiation based on CNVR

The normalized copy number on CNVRs of all individuals was calculated using CNVnator (Abyzov et al., 2011). V_{ST} of normalized copy number between a pair of breeds, was calculated as $V_{ST} = (V_T - V_S) / V_T$ where V_T is the total variance of normalized copy number among all individuals from both breeds and Vs is the average of variance within each breed, weighted by the number of individuals in the breed (Redon et al., 2006). After excluding the 6 breeds with single individual, V_{ST} between pairs of 33 *bos taurus* breeds and a buffalo breed were calculated. Mean V_{ST} of all autosomal CNVRs in each pair of breeds were visualized using pheatmap in R (Kolde, 2012). In addition, the V_{ST} of autosomal CNVRs were calculated between EAT, ASI, AFH and AFT. These results were visualized as Manhattan plots using qqman package in R (Turner, 2014). After ranking the normalized copy numbers of all B. taurus individuals, *Kruskal-Wallis* test implemented in 'kruskal.test' R function were performed on all autosomal CNVRs to compare populations inlcuding EAT, ASI, AFH and AFT. Population differentiated CNVRs were defined as autosomal CNVRs with top 1% pairwise as well as *Kruskal-Wallis* test p-value less than 0.01.

2.3.6. Functional annotation of genes overlapped with candidate CNVRs

Genes overlapped with autosomal CNVRs were annotated based on the reference genome ARS-UCD1.2 from NCBI RefSeq database (O'Leary et al., 2016). In case of genes overlapped with multiple CNVRs, the CNVR with the most significantly different in *Kruskal-Wallis* test was written. Hypothetical, putative, predicted, or uncharacterized genes and pseudo-genes were excluded. The information of functional annotation, gene ontology and pathway of the genes within the population differentiated CNVRs were identified using PANTHER classification system (Mi et al., 2019). Comparing the list of genes overlapped with CNVRs with the all genes of *bos taurus* in PATHER database (Mi et al., 2013), I tested the hypothesis whether the PANTHER GO-slim molecular function, GO-slim biological process, and pathway terms were under- or overrepresented in CNVRs using binomial test with Bonferroni corrections (Mi et al., 2019; Nicholas et al., 2009). The quantitative trait loci (QTL) underlying CNVRs were also identified using Cattle QTLdb of the reference genome ARS-UCD1.2 (Hu et al., 2019). Under- or overrepresentation of autosomal QTL in autosomal CNVRs was tested using binomial test with *Bonferroni* corrections.

2.4. Results

2.4.1. CNV calling and CNVR definition

The coverage and sequencing depth of mapped short reads data are important to reliably call CNVs using read depth information. In several previous studies, samples with mean depth coverage over 5x were used for CNV analysis, showing that 4x depth coverage is sufficient for read depth-based CNV detection (Bickhart et al., 2012; Consortium, 2012; Sudmant et al., 2010). In my dataset, the minimum mean depth was higher than 5.1x, and the mean values of alignment rate, coverage and mean depth of coverage were 99.5%, 95.0%, 11.4x (Table 2.1). After calling and filtering CNVs, 18391 CNVRs were identified on autosomes, covering 236.2 Mbp or 9.49% of B. taurus autosomes.

2.4.2. Population differentiation based on CNVR

In the hierarchical clustering tree based on CNVR, 8 individuals including a Maremmana (MAM03), a Maronesa (MAN01), 4 Jersey individuals (JER03, JER04, JER05 and JER06), an Angus (ANG09) and an Ankole (ANK03) were distant from other individuals (Figure 2.1). Except for the 8 individuals, 330 individuals which consisted of 2 AFB, 211 ASI or AFH (indicine group), 117 EAT or AFT (taurine group) were classified by their species and subspecies. Most of the taurine individuals were clustered by their breeds in contrast to indicine individuals. The AFT individuals were grouped by their breeds and were separated from EAT breeds that were mostly well clustered by their breeds. The four EAT breeds, Holstein, Hanwoo, Hereford and Simmental, were distinguished from other breeds and all individuals in each breed were grouped together.

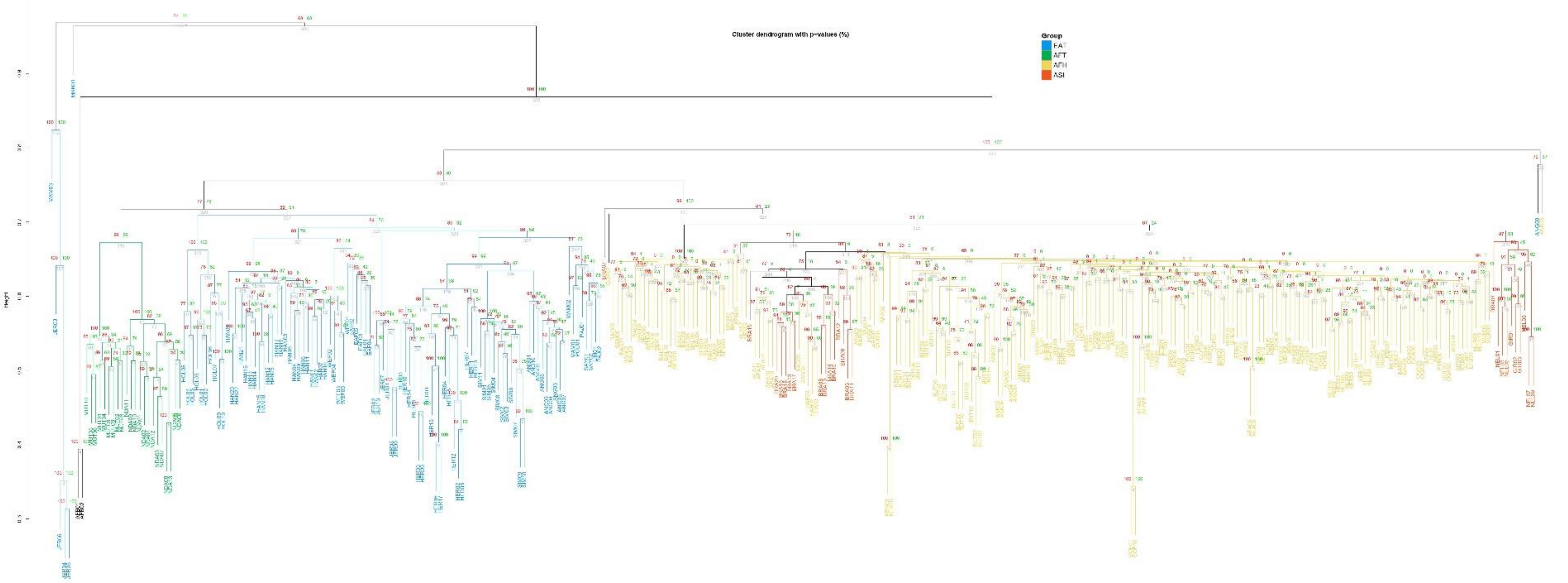

Figure 2.1. Hierarchical clustering tree.

For every individual, the absence or presence of CNVs in autosomal CNVRs was converted to vector made of ‘0’s and ‘1’s. The hierarchical clustering was performed on these vectors representing each individual. The bootstrap value was written under the edges of every clustering. The approximately unbiased (AU) and the bootstrap probability (BP) p-value were written in red and green letters on the edges after being multiplied by 100. The branch of hierarchical clustering tree were colored to indicate the group of clades following their region and population such as AFB, AFH, AFT, ASI and EAT.

The individuals of two Finn cattle breeds, Western Finn and Eastern Finn, were not distinguished from each other, but clustered together. 6 of 10 Angus and 9 of 10 Jersey individuals were clustered and differentiated by their breeds. Rest of the taurine individuals included in Maremmana, Podolica, Pajuna, Sayaguesa and Limia from South-Western Europe were grouped together. While Nelore and Gir were distinguished from AFH, individuals in other ASI breeds such as Brahman, Sahiwal, Tharparkar and Hariana were clustered with AFH individuals.

The variance of copy numbers of each breed and V_{ST} of breed pairs were calculated for every autosomal CNVR. The range of V_{ST} is from 0 to 1, with a higher value indicating a larger difference. The pairwise mean V_{ST} of regional population were as following: EAT-AFT, 0.008; EAT-ASI, 0.017; AFH-ASI, 0.023; AFH-EAT, 0.024; AFH-AFT, 0.045; AFT-ASI, 0.128 (Figure 2.2). The average of the mean of pairwise V_ST in breed level was 0.166. Most of the AFH and ASI were clustered together and N'Dama was clustered with EAT. Muturu was clustered with the 3 Ethiopian humped breeds including Bagaria, Bale and Semien, and separated from others. Several groups of breeds originated from adjacent region including Finn taurine (Eastern Finn and Western Finn), and the Ethiopian zebu (Bagaria, Bale and Semien) were clustered together by their mean V_{ST}.

2.4.3. Detection of candidate CNVR differentiated across populations

In order to detect population differentiated CNVR across 4 groups (AFH, AFT, ASI, and EAT), two statistics were employed. First, pairwise V_{ST} were calculated between all populations except for AFB. Top 1% and top 0.1% values were about 0.500 and 0.759, respectively. The number of CNVRs with the top 0.1% V_{ST} was 109 in ASI-AFT pair, 2 in ASI-EAT pair and 0 in other pairs.

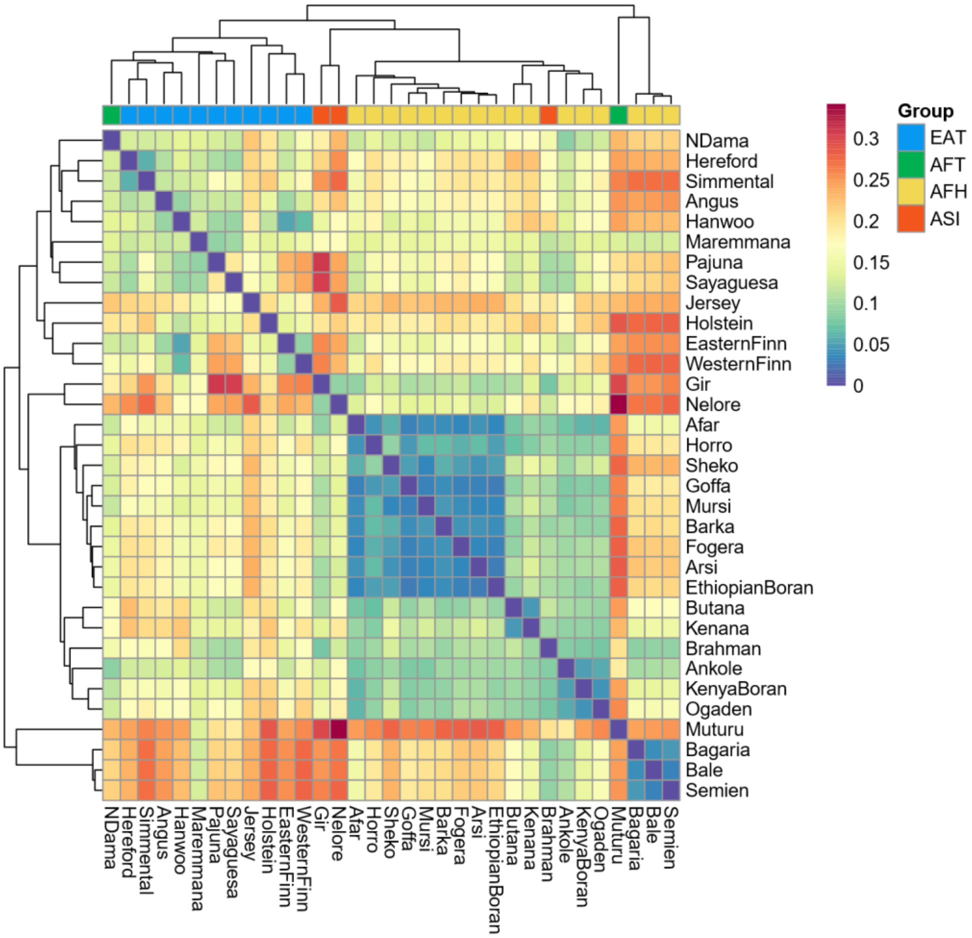

Figure 2.2. Heatmap of Mean pairwise V_{ST} values between cattle breeds represented by more than one animal.

Clustering tree and heatmap of mean pairwise V_{ST} of autosomal CNVRs. The group of breeds was visualized by color above each column. The arrangement of breeds in row and column followed the order by clustering tree. The agglomeration method of clustering was weighted pair group method with arithmetic mean (WPGMA). Breeds were classified to 4 groups by their originated region and taxonomy as follows; AFH, African Humped cattle; AFT, African humpless taurine; ASI, Asian indicus; EAT, Eurasian taurine.

The number of CNVRs with a higher V_{ST} than top 1% pairs of populations as follows: 1033 in ASI-AFT pair, 31 in EAT-ASI pair, 21 in EAT-AFH pair, 15 in AFH-AFT pair and 2 in both ASI-AFH pair and EAT-AFT pair. The V_{ST} of pairs of 4 regional B. taurus populations: EAT, ASI, AFT and AFH were visualized as Manhattan plots (Figure 2.3). Then, differences in rank of normalized copy number across 4 groups of B. taurus including ASI, EAT, AFT and AFH were tested using *Kruskal-Wallis* test. The population differentiation of CNVRs were determined by the following two criteria: p-value under 0.01 in *Kruskal-Wallis* test and pairwise V_{ST} in upper 1% which resulted in 910 CNVRs including 313 genes as candidates.

2.4.4. Functional annotation of CNVR overlapping genes

Among 313 genes overlapped with 362 of population differentiated CNVRs, those with average copy number which is different between populations are summarized in Table 2.2. The differentiated CNVRs were sorted in ascending order of chi-square from *Kruskal-Wallis* test. The average copy numbers for AFT, AFH, ASI, EAT groups were written under column for each group. Significantly under- or overrepresented PANTHER GO-slim molecular functions, GO-slim biological processes, or pathways were summarized in Table 2.3. Most of GO terms with significantly different representation between CNVRs and genome were overrepresented. Regulation associated terms including RNA polymerase II specific DNA binding, DNA-binding transcription factor, regulation of transcription by RNA polymerase II were overrepresented in CNVRs. Nervous system development and cell differentiation related terms were overrepresented, while immune response and structural constituent of ribosome were underrepresented.

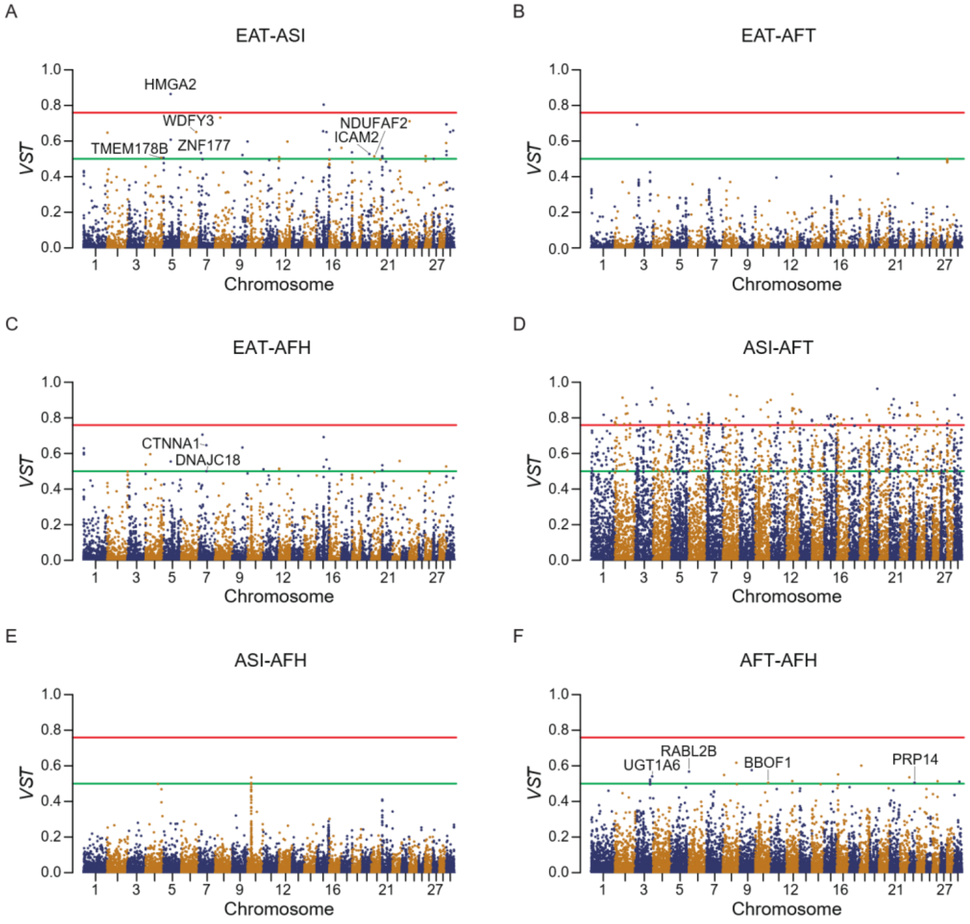

Figure 2.3. Manhattan plot of V_{ST}.

V_{ST} of CNVRs were visualized as Manhattan plots. The center point of CNVRs was used as x-coordinate value. Differentiated genes overlapped with CNVRs significantly different both in upper 1% V_{ST} and 0.01 significance level of *Kruskal-Wallis* test on their copy number. The genes whose symbol is starting with 'LOC' or differentiated in ASI-AFT pair were left out due to lack of space. The upper 1% percentile V_{ST}, 0.500 and upper 0.1% percentile, 0.759 were shown as green and red lines respectively.

Table 2.2 Genes overlapped with population differentiated CNVRs.

Genes overlapped with significantly different CNVRs based on Kruskal-Wallis test result with <0.01 significance level and upper 1% VST. Genes on CNVRs were sorted in ascending order by p-values. The pairs of populations with top 1% or top 0.1% V_{ST} and the average of copy number of CNVRs in populations including EAT, AFT, AFH and ASI are also indicated.

CNVR	**Chr.**	**Start**	**End**	**p-value**	**Gene List**	**EAT**	**AFT**	**AFH**	**ASI**
7993_DUP	10	79275201	79278200	2.20E-16	EIF2S1	3.56	4.00	2.61	2.35
3638_DEL	5	58027201	58090800	2.20E-16	OR6C202	1.91	1.49	0.58	0.48
14628_DUP	21	28806801	28824600	2.20E-16	TM2D3	2.39	3.21	4.95	5.46
5686_DEL	7	50070401	50072400	2.20E-16	CTNNA1	1.90	1.38	0.36	0.44
11438_DUP	15	79702801	79724600	2.20E-16	OR8U3	2.11	2.34	2.91	3.30
10728_DUP	15	628601	641800	2.20E-16	OR4C1N	2.08	2.31	3.14	3.23
13749_DUP	19	41438001	41471600	2.20E-16	KRTAP9-1, KRTAP9-2	3.92	2.73	2.00	1.91
11581_DUP	15	84704401	84712000	2.20E-16	OR4C181	2.15	2.33	3.24	3.28
11923_DUP	16	53879001	53881800	2.20E-16	PRDM2	2.21	2.34	3.71	3.32
11580_DUP	15	84704001	84729200	2.20E-16	OR4C181	2.08	2.18	2.83	2.78
11440_DUP	15	79716001	79724800	2.20E-16	OR8U3	2.11	2.37	3.23	3.90
11569_DUP	15	84423401	84427400	2.20E-16	OR4A16	2.16	2.44	3.26	3.50
5690_DEL	7	50689601	50691400	2.20E-16	DNAJC18	1.96	1.63	0.31	0.26
2299_DEL	3	113312401	113317200	2.20E-16	UGT1A6	1.98	1.88	0.75	0.34

4710_DUP	6	72332401	72335600	2.20E-16	RESTA	2.11	2.21	3.43	2.88
18250_DUP	29	44417801	44435200	2.20E-16	SLC29A2	2.11	2.16	2.76	2.63
15188_DUP	22	51590001	51603400	2.20E-16	CATHL4	2.20	2.01	3.70	3.76
13961_DEL	19	62786201	62788800	2.20E-16	PRKCA	2.24	4.62	1.33	0.35
13750_DUP	19	41439001	41442600	2.20E-16	KRTAP9-2	4.43	3.09	2.48	2.36
13803_DUP	19	48209801	48215000	2.20E-16	ICAM2	2.22	2.29	4.06	4.17
8144_DEL	10	102401201	102404000	2.20E-16	TTC7B	1.97	2.01	1.37	0.97
8442_DEL	11	55496001	55499400	2.20E-16	CTNNA2	0.69	0.85	1.67	1.81
3027_DUP	4	105937801	105942200	2.20E-16	TCRB	3.51	3.45	2.47	1.99
10980_DEL	15	43525601	43527800	2.20E-16	SCUBE2	1.94	1.98	1.14	0.42
11579_DUP	15	84693601	84712000	2.20E-16	OR4C181	2.10	2.15	2.74	2.73
5400_DUP	7	15954401	15980200	2.20E-16	HNRNPA2B1	2.03	1.96	2.85	2.51
3560_DUP	5	47840001	47846200	2.20E-16	HMGA2	2.37	2.48	5.13	8.85
2971_DEL	4	104569001	104575600	2.20E-16	TMEM178B	2.00	2.07	0.99	0.69
1193_DEL	2	61661801	61663200	2.20E-16	R3HDM1	0.54	0.56	1.79	1.85
14629_DUP	21	28818601	28824600	2.20E-16	TM2D3	2.83	4.90	9.03	10.44
5049_DUP	7	305401	343400	2.20E-16	OR5W39	2.14	2.17	2.71	2.58
10934_DEL	15	29761201	29764000	2.20E-16	NLRX1	1.99	1.83	0.98	0.50
16822_DEL	26	7143601	7149600	2.20E-16	PRKG1	1.95	2.11	1.34	0.77
4280_DUP	5	120030201	120081800	2.20E-16	RABL2B	6.89	7.62	4.63	3.44
6021_DEL	8	317001	330000	2.20E-16	MFSD14B	1.89	1.64	0.93	0.45

3162_DEL	4	114207401	114229400	2.20E-16	PRKAG2	0.78	0.64	1.08	1.00
3559_DUP	5	47822601	47856200	2.20E-16	HMGA2	2.08	2.21	3.13	3.31
15424_DEL	23	15094601	15097200	2.20E-16	TREM2	0.58	0.93	1.63	1.88
3026_DUP	4	105935201	105949600	2.20E-16	TCRB	3.22	3.22	2.31	1.92
9655_DUP	13	2825001	2836800	2.20E-16	PAK5	2.51	2.84	4.40	4.26
12720_DEL	18	8632001	8638000	2.20E-16	HSD17B2	1.94	1.66	1.50	1.09
8466_DEL	11	60510601	60514000	2.20E-16	COMMD1	2.00	1.94	1.14	0.97
13751_DEL	19	41443801	41447800	2.20E-16	KRTAP9-2	4.11	2.76	1.89	1.79
14983_DUP	22	1614001	1618200	2.20E-16	NEK10	2.14	1.99	2.75	2.83
8416_DEL	11	50478201	50483200	2.20E-16	DNAH6	2.12	1.95	1.29	1.19
7155_DEL	9	86997401	86999800	2.20E-16	LRP11	0.76	0.26	1.24	1.80
1090_DEL	2	30769401	30770800	2.20E-16	CSRNP3	1.94	1.95	0.88	0.87
15201_DUP	22	53204001	53226200	2.20E-16	CCR1	2.07	2.09	2.77	3.29
17767_DEL	28	27820801	27823000	2.20E-16	CDH23	0.85	1.12	1.82	2.28

Table 2.3. Over- / underrepresentation of PANTHER GO-slim molecular function, GO-slim biological process and pathway terms on CNVRs.

	Bos taurus	**CNVRs**	**Expected**	**Over/Under representation**	**P-value**
Molecular Function	**22704**	**6297**			
RNA polymerase II regulatory region sequence-specific DNA binding	346	168	96.0	+	5.9E-09
transcription regulatory region sequence-specific DNA binding	377	181	104.6	+	2.4E-09
transcription regulatory region DNA binding	432	202	119.8	+	1.4E-09
regulatory region nucleic acid binding	432	202	119.8	+	1.4E-09
sequence-specific double-stranded DNA binding	400	182	110.9	+	1.3E-07
sequence-specific DNA binding	529	216	146.7	+	1.7E-05
double-stranded DNA binding	441	193	122.3	+	7.2E-07
RNA polymerase II regulatory region DNA binding	351	172	97.4	+	1.7E-09
DNA-binding transcription factor activity, RNA polymerase II-specific	372	172	103.2	+	1.3E-07
DNA-binding transcription factor activity	608	270	168.6	+	8.8E-11
transcription regulator activity	767	335	212.7	+	8.8E-13
structural constituent of ribosome	259	21	71.8	-	6.5E-10
structural molecule activity	360	52	99.9	-	4.0E-05
Biological Process					
positive regulation of transcription by RNA polymerase II	361	175	100.1	+	9.0E-09
regulation of transcription by RNA polymerase II	779	339	216.1	+	3.4E-12
regulation of transcription, DNA-templated	1102	434	305.6	+	1.2E-09
regulation of cellular macromolecule biosynthetic process	1205	459	334.2	+	2.7E-08
regulation of cellular biosynthetic process	1224	463	339.5	+	5.6E-08

regulation of biosynthetic process	1231	464	341.4	+	8.5E-08
regulation of metabolic process	1987	683	551.1	+	1.2E-05
regulation of biological process	4183	1341	1160.2	+	7.4E-06
biological regulation	4655	1502	1291.1	+	9.7E-08
regulation of cellular metabolic process	1853	649	513.9	+	1.9E-06
regulation of cellular process	3947	1273	1094.7	+	5.7E-06
regulation of macromolecule biosynthetic process	1212	459	336.2	+	5.8E-08
regulation of macromolecule metabolic process	1891	664	524.5	+	8.1E-07
regulation of nucleic acid-templated transcription	1102	434	305.6	+	1.2E-09
regulation of RNA biosynthetic process	1102	434	305.6	+	1.2E-09
regulation of RNA metabolic process	1217	467	337.5	+	6.4E-09
regulation of nucleobase-containing compound metabolic process	1263	479	350.3	+	1.7E-08
regulation of nitrogen compound metabolic process	1791	638	496.7	+	2.1E-07
regulation of primary metabolic process	1815	641	503.4	+	7.1E-07
regulation of gene expression	1357	512	376.4	+	6.8E-09
positive regulation of transcription, DNA-templated	446	203	123.7	+	4.2E-08
positive regulation of nucleic acid-templated transcription	446	203	123.7	+	4.2E-08
positive regulation of RNA biosynthetic process	446	203	123.7	+	4.2E-08
positive regulation of RNA metabolic process	460	208	127.6	+	4.2E-08
positive regulation of macromolecule metabolic process	833	321	231.0	+	1.1E-05
positive regulation of metabolic process	856	327	237.4	+	1.9E-05
positive regulation of nucleobase-containing compound metabolic process	479	214	132.9	+	6.0E-08
positive regulation of cellular metabolic process	804	315	223.0	+	3.2E-06
positive regulation of cellular process	1226	440	340.0	+	8.6E-05
positive regulation of nitrogen compound metabolic process	805	318	223.3	+	1.2E-06
positive regulation of macromolecule biosynthetic process	480	209	133.1	+	8.2E-07
positive regulation of biosynthetic process	491	211	136.2	+	1.9E-06

positive regulation of cellular biosynthetic process	488	211	135.4	+	1.2E-06
positive regulation of gene expression	506	213	140.3	+	7.9E-06
transcription by RNA polymerase II	809	348	224.4	+	6.5E-12
transcription, DNA-templated	1145	446	317.6	+	2.7E-09
nucleic acid-templated transcription	1145	446	317.6	+	2.7E-09
RNA biosynthetic process	1151	447	319.2	+	3.8E-09
cellular macromolecule metabolic process	3268	1061	906.4	+	5.2E-05
anatomical structure development	1034	399	286.8	+	1.3E-07
developmental process	1168	451	324.0	+	6.6E-09
multicellular organism development	821	323	227.7	+	1.3E-06
cell differentiation	733	289	203.3	+	8.4E-06
cellular developmental process	800	311	221.9	+	8.5E-06

Among 72,840 of autosomal QTLs, 7,699 of QTLs overlapped with CNVR. 5,252 of QTLs overlapped with duplication CNVR and 2,642 of QTLs overlapped with deletion CNVRs. The representation of QTLs related to reproduction, milk and body weight were significantly different compared to total QTL. In reproduction related QTLs, the luteal activity was underrepresented on CNVRs while non-return rate, gestation length and calving ease were overrepresented. Most luteal activity QTL overlapping CNVRs were duplication while most gestation length QTL were overlapped with deletion. The milk content related QTLs such as milk kappa-casein, glycosylated kappa-casein, unglycosylated kappa-casein percentage and milk potassium content were underrepresented on CNVRs. On the other hand, milk fat and yield QTLs were overrepresented. Body weight (yearling) and body weight gain QTLs were underrepresented on CNVRs.

2.5. Discussion

Cattle have been spread with humans across the world after the domestication event in the Fertile Crescent in 10,000 YBP and Indus Valley in 8,000 YBP. The genetic fenvironments and demographic history including migration and introgression. For example, the population structure of African cattle has diversely changed from its earliest taurine-like population. Since the arrival of *B. indicus* around 700 AD (Hanotte et al., 2002; Stock & Gifford-Gonzalez, 2013), the taurine × indicine cattle admixture event 750-1,050yr ago (Kim et al., 2020) and the introgression of African aurochs constructed the complex population structure of the current African cattle. Although population genetics of cattle has been studied extensively based on SNPs, the effects of CNVs on phenotypes and signatures of evolution were poorly understood.

CNVs cover a larger region of genome than SNPs and can impact gene function in multiple ways, including changing of gene structure and dosage, altering gene regulation and exposing recessive alleles (Zhang et al., 2009). Notably, genes overlapping CNVs were shown to have better correlations with differentially expressed genes than nearby SNPs, particularly when the CNV overlapped with exons (Schlattl et al., 2011). Deletions in cattle genome can impact phenotype by interrupting genes and causing loss of biological function (Liu & Bickhart, 2012). Duplicated genes in cattle genome were related to digestion, lactation, reproduction and immune system such as antigen processing and major histocompatibility genes (Keel et al., 2016; Liu et al., 2009). CNVs also have population genetic nature related to recombination, mutation, selection, and demography (Sjödin & Jakobsson, 2012). Generally, CNVs are more recent events than SNPs as they are still segregating within population, showing greater inter-individual variability (Mielczarek et al., 2018). These functional impacts and population genetic nature of CNVs have suggested that population differentiation of CNVs may contribute to the phenotypic variation between populations.

Recently, high quality cattle genome assemblies such as ARS-UCD1.2, UOA_Angus_1 and UOA_Brahman_1 increased reliability of CNV calling and resolution of breakpoint. Above all, Low et al. released haplotype-resolved genome assemblies of of *bos taurus taurus* and *bos taurus indicus*, and compared CNV between two subspecies (Low et al., 2020). They performed CNV calling using short reads from 38 animals of 7 cattle breeds.

I expanded samples to 336 individuals in 39 global cattle breeds in present study. I aligned short reads on ARS-UCD1.2 assembly to compare larger populations under unified criteria. I identified population stratification of autosome-wide CNVs based on NGS read mapping. Particularly, I included 206 individuals of 19 African cattle breeds in which their genome-wide CNV have been analyzed for the first time in this study.

The traditional classification for African indigenous cattle was based on phenotypes, especially the existence of cervico-thoracic hump. Based on this, some of the hybridized breeds were called Sanga (Zebu x Taurine) and Zenga (Zebu x Sanga). However, genome-wide SNP analysis has identified that the traditional classification did not reflect the genetic difference well (Bahbahani et al., 2018; Edea et al., 2015). My CNV based classification generally agreed with previous knowledge with exceptions in several individuals. There were two reasons for the disagreement. Firstly, this study only covered copy number variation region, not the entire genome. Secondly, I compared the read mapping-based copy number, not the sequence itself.

Nevertheless, overall concordance of clustering showed potential for population stratification using CNV.

In my CNV-based hierarchical clustering, most individuals were classified by their breeds, whereas some individuals including MAM01, MAM03, ANG09, ANK03 and part of Jersey individuals separated from their breeds. I inspected two possibilities to figure out the reason of the inconsistency. First, I checked similarity between individuals in each breed. I referred to my previous study sharing large part of dataset (Kim et al., 2020). The PCA plot and population structure from SNP genotype indirectly verified that there were no individuals significantly distinguished from their breeds. Second, the input vector of hierarchical clustering was the next suspicious one after excluding sample problem. It was too simple to represent CNV enough. The element of vector only considered existence of CNV on each CNVR, neglecting other properties such as length, breakpoint and copy number of CNV. I also tested two other vectors indicating type of CNV and normalized copy number of CNV. But the vector considering existence of CNV on CNVR made hierarchical tree which was the most concordant with breeds. Third, greater inter-individual variability of CNVs compared to SNPs and indels might contributed to this discordance (Mielczarek et al., 2018).

Mean V_{ST} and the number of CNVRs with high V_{ST} supported the ancestry of African cattle. AFT-EAT and AFH-ASI pairs were relatively similar while the AFT-ASI pair was mostly different. AFH exhibited high levels of shared CNV with ASI but not with AFT, probably because of their recent admixture around 150 generations ago (Kim et al., 2020). Pairwise comparison of breed distinguished Muturu from others, and clustered with the 3 Ethiopian zebu; Bagaria, Bale and Semien. The African taurine, especially Muturu, showed no evidence of admixture in previous studies assuming EAT and Asian-Australian indicine (AAI) as proxies for unadmixed taurine and indicine cattle, respectively (Kim et al., 2020). Muturu was separated from EAT, ASI, and most of AFH except for Bale, Bagaria and Semien in pairwise mean V_{ST} clustering tree. Although the 3 Ethiopian breeds were clustered with Muturu, the mean pairwise V_{ST} did not imply their closeness to Muturu. The mean V_{ST} of Bale, Bagaria and Semien were 0.249, 0.244 and 0.251, respectively, which were all similar with the average 0.249. In addition, Italian taurine, Maremmana (0.132) and the Iberian indigenous taurine, Sayaguesa (0.189) and Pajuna (0.199) have lowest mean V_{ST} against Muturu, which supported the shared ancestry between Muturu and Southern European taurine (Kim et al., 2020; Upadhyay et al., 2019).

Based on the V_{ST} and *Kruskal-Wallis* test on the copy number of CNVRs, 313 genes were obtained as candidate genes under selection and adaptation. Of those, several genes were related to disease susceptibility and resistance. I identified significantly higher copy number of *HMGA2* in indicine than in taurine. The indicine-specific copy number gain of *HMGA2* was identified by chip-based methods and validated using qPCR in a previous study in which the *HMGA2* duplication in Nellore was suggested to be associated with navel length at yearling by haplotype-based GWAS (p = 1.01 × 10^{-9}) (Aguiar et al., 2018). Navel length at yearling is an economically important trait related to navel injuries in beef cattle. A pendulous navel increases the risk of injuries and infection caused by friction against the pasture (Rabelo et al., 2008). In natural mating, bulls with long and pendulous navels would be frequently exposed to injuries and trauma (Boligon et al., 2016). Expression of *HMGA2* gene is also responsible for body size by regulating myoblast proliferation and myogenesis. *HMGA2* directly regulates transcription of *IGF2BP2* (insulin like growth factor 2 mRNA binding protein 2), and *IGF2BP2* promotes myoblast growth. *IGF2BP2* regulates translation of *IGF1R* (insulin like growth factor 1 receptor), *c-Myc*, and/or *Sp1* by binding to their mRNA (Z. Li et al., 2012). Among these genes related to muscle growth, *HMGA2*, *IGF2BP2* and *IGF1R* were overlapped with my CNVRs. The copy number of overlapping CNVRs of *HMGA2* and *IGF1R* was significantly different between populations whereas *IGF2BP2* overlapping CNVR was not. The copy number of *HMGA2* overlapping CNVR was gained in indicine population (EAT: 2.37, AFT: 2.48, AFH: 5.13, ASI: 8.85). On the contrary, the copy number of *IGF1R* overlapping CNVR was gained in taurine population and lost in indicine population (EAT: 3.28, AFT: 4.34, AFH: 0.92, ASI: 0.43). The knockout mice experiment suggested the positive impact of *HMGA2* expression on myoblast growth (Z. Li et al., 2012). On the other hand, Chinese beef cattles with copy number loss of *IGF1R* had significantly better growth trait such as body weight, body height and hucklebone width (Ma et al., 2019). In addition, *HMGA2* and *IGF1R* were also strongly associated with size differences between dog breeds (Sutter et al., 2007). In conclusion, I suggest that differentiated copy number of *HMGA2* and *IGF1R* might contribute to make differences in body size between populations. Copy number variable genes overlapped with taurine-specific duplication such as *KRTAP9-1* and *KRTAP9-2*, and indicine-specific duplication such as *CATHL4* and *PRDM2* are related to pathogen- and parasite-resistance. The taurine-specific duplication of *KRTAP9-1* and *KRTAP9-2* corroborates the previous result of comparing copy number of them between European taurine and Asian zebu (Bickhart et al., 2012; Bickhart et al., 2016). They were also identified by aligning WGS short reads to three reference genome assemblies including ARS-UCD1.2, UOA_Angus_1 and UOA_Brahman_1 (Low et al., 2020). The keratin associated proteins were suggested to play a role in tick resistance (Nakamura et al., 2013; Wang et al., 2007). Since the cattle skin is the infestation site of tick, the structural protein keratin which makes up the outer layer of skin and hair could act as a barrier (Taye et al., 2018). Also, the *PRDM2* gene was referred to play a role in resistance to disease and bacterial infection or cell-mediated immune response, especially paratuberculosis resistance in ruminants (Ghoreishifar et al., 2020; Moioli et al., 2016). The Paratuberculosis (Johne's disease) caused by *Mycobacterium avium* subspecies *paratuberculosis* (*MAP*) brought about considerable economic losses worldwide. The GWAS cohort study about MAP infection in Holstein cattle identified strong signal of SNP and QTL adjacent to *PRDM2* gene (Mallikarjunappa et al., 2018). Although the resistance to *MAP* has not yet been compared between taurine and indicine, the *PRDM2* gene overlapping indicine-specific duplication in my result can be the candidate region for further investigation on adaptation and selection related to paratuberculosis. The higher copy number of *CATHL4* in ASI than EUT was also identified in a previous study (Bickhart et al., 2012). The bovine reference genome contains the expanded antimicrobial cathelicidine gene family whereas humans and mice have single copy (Elsik et al., 2009). Especially, the antimicrobial peptide, indolicidine encoded by *CATHL4* can induce autophagic cell death of *Leishmana donovani*, which is the causative parasite of Leishmaniasis (Bera et al., 2003). The antimicrobial ability which can influence Leishmaniasis lesion development of CATH-family genes was also proved by a knockout in mice (Kulkarni et al., 2011). Taken together, the population differentiated CNV on these genes may contribute to the increased parasite resistance in indicine compared to taurine.

ASI found across the tropical Indian subcontinent adapted to tropical environments characterized with heat stress as well as pervasive pathogen such as tick and parasite (Chan et al., 2010). AFH whose ancestry of selection signature skewed toward indicine was also suggested to be adapted to heat stress by indicine introgression into local taurine (Kim et al., 2020). In my analyses, one of the heat shock protein family coding gene, *DNAJC18* is found to be overlapped with indicine-specific deletion, which is consistent with the CNVR identified in a previous study (Hu et al., 2020). The DnaJ family binds to HSP70s for regulating their client capture and drives HSP70s toward specific client (Kampinga & Craig, 2010). The significantly higher contribution of indicine ancestry (Kasarapu et al., 2017) and selection signature in East African short horn zebu (Bahbahani et al., 2017) imply that CNV on *DNAJC18* play a role in tropical adaptation and heat tolerance of zebu.

The olfactory function has evolved to alert animals to presence of possible threats such as predators, and provides ability to avoid dangerous food containing harmful parasites, bacteria or chemicals (Reed & Knaapila, 2010). It also assists animals in locating foods and potential mates. (Spehr & Munger, 2009). Olfactory receptors (ORs) play the key role in olfactory function, detecting odor molecules in the olfactory epithelium of the nasal cavity. The OR genes are the largest gene family in the mammalian genome, and there are 881 OR genes in cattle (Lee et al., 2013). The OR genes are also characterized by extremely frequent gene duplications and losses (Niimura, 2012). In cattle, about 40% of OR loci are identified as CNVs. Therefore, the diversity and CNVs on OR genes in cattle could lead to breed specific differences in olfaction capacity (Lee et al., 2013). In my result, several OR genes were overlapped with the population differentiated CNVRs. There were *OR6C202*, *OR10AD1* and *OR5T2* on indicine-specific deletion, *OR8U3*, *OR4C1N*, *OR4C181*, *OR2AP1*, *OR9K2*, *OR4A16* and *OR5D14* on indicine-specific duplication, *OR4S1*, *OR5T2*, *OR8K1* and *OR5AS1* on ASI-specific deletion, *OR5M3* and *OR5AR1* on ASI-specific duplication and *OR8K3*, *OR5AS1* and *OR5L2* on African cattle specific duplication. As the significant variations in the number and repertoires of OR gene among vertebrates indicate that olfactory function has strongly influenced by natural selection my specific set of OR CNVs might give candidate CNVRs under selection.

Copy numbers of genes associated with quantitative traits related to productivity were frequently gained or lost on cattle genome. In my results, the Eukaryotic translation initiation factor 2 subunit 1 (*EIF2S1*) gene was overlapped with taurine-specific duplication from 7927275.2 to 79278.2 kb in chromosome 10. Copy number on the CNVR in ASI-AFT pair was significant in one-way ANOVA test and their V_{ST} was 0.887. In previous study, *EIF2S1* was overlapped with CNVR specific to a high feed efficient group of Holstein (Hou et al., 2012), which suggests the contribution of the CNVR to different feed efficiency in beef cattle between *bos taurus taurus* and *bos taurus indicus* (Canal et al., 2020; Sainz et al., 2013). The muscle development related gene *CTNNA1* was overlapped with indicine-specific deletion. This result was mostly agreed by Hu et al. (Hu et al., 2020) except for the lower copy number in my AFT individuals. The low copy number in *bos taurus indicus* while normal or little change in *bos taurus taurus* suggest that the sequence is likely to be specific to *bos taurus taurus*. The *CTNNA1* gene has been described to be associated with myostatin expression level and transcription in skeletal muscle in Holstein-Friesian bulls (Sadkowski et al., 2008). Since myostatin plays an essential role in regulating skeletal muscle growth, the taurine-specific existence of *CTNNA1* gene would be one of the explanations for difference in meat productivity between *bos taurus taurus* and *bos taurus indicus*.

This chapter was published in *Scientific Reports*

as a partial fulfillment of Jisung Jang's Ph.D program,

Chapter 3. Population differentiated copy number variation between Eurasian wild boar and domesticated pig populations

3.1. Abstract

Sus scrofa is a globally distributed livestock species that still maintains two different ways of life: wild and domesticated. Herein, I detected copy number variation (CNV) of 328 animals using short read alignment on Sscrofa11.1. I compared CNV among five groups of porcine populations: Asian domesticated (AD), European domesticated (ED), Asian wild (AW), European wild (EW), and Near Eastern wild (NEW).

In total, 21,673 genes were identified on 154,872 copy number variation region (CNVR). Differences in gene copy numbers between populations were measured by considering the variance-based value V_{ST} and the one-way ANOVA test followed by *Scheffe* test. As a result, 111 genes were suggested as copy number variable genes. Abnormally gained copy number on *EEA1* in all populations was suggested the presence of minor CNV in the reference genome assembly, Sscrofa11.1. Copy number variable genes were related to meat quality, immune response, and reproduction traits. Hierarchical clustering of all individuals and mean pairwise V_{ST} in breed level were visualized genetic relationship of 328 individuals and 56 populations separately. My findings have shown how the complex history of pig evolution appears in genome-wide CNV of various populations with different regions and lifestyles.

3.2. Introduction

Pig (*Sus scrofa*) is by far one of the most globally distributed animal species maintaining two different ways of life: wild and domesticated. The great adaptability of wild boar makes it possible to colonize the wild areas, including mainland Eurasia and North Africa, within 2 Mya, after originating from Southeast Asia in the early Pliocene 5.3–3.5 Myr ago (Groenen et al., 2012). In addition to adaptation to various environments of the wide habitats, demographic events such as migration and bottleneck during the glacial periods also make pigs diverge into numerous populations. The two main populations of wild boar, European and Asian, diverged around 1 Mya (Frantz et al., 2013). Initial domestication took place independently at two locations, East Anatolia and China with local wild boars in 9,000 to 10,000 years ago (Fang et al., 2009). Mitochondrial DNA analysis by (mtDNA) suggested that European domesticated pigs arrived from Near East alongside farmers 8,500 YBP (Frantz et al., 2019).

The population and geographical distribution of the domesticated pigs have greatly varied from wild boars after initial domestication because of long-term climate fluctuations, human hunting, and follow-up stock-raising activities (Larson et al., 2010). However, domesticated pigs and wild boars were not only consistently diverged from one another. For instance, over 3,000 years after the arrival of Near Eastern domesticated pigs to Europe, domesticated pigs were interbred with local wild boar. It made most of Near Eastern ancestry disappear in the genomes of European domesticated pigs (Frantz et al., 2019; Paudel et al., 2015). Subsequent selection and breeding of domesticated pigs resulted in highly distinct pig breeds in Europe and Asia (Megens et al., 2008). Domesticated pigs have undergone a complex history of selection and migration to improve commercial traits. For example, European farmers induced introgression between Asian and European domesticated pigs to improve commercial traits such as litter size and backfat in the early nineteenth century (White, 2011). Modern breeding practices, including reproductive isolation and genomic selection, have accelerated genetic divergence between wild boar and domesticated pigs since the foundation of modern pig breeds, starting around 200 years ago. Previous genome-wide SNP studies identified distinct patterns of selection in domesticated pigs and wild boars (M. Li et al., 2013; Wilkinson et al., 2013).

Copy number variation (CNV) is another type of variation which covers more significant part of the porcine genome than single nucleotide polymorphism (SNP). CNV can be a major mechanism driving genome evolution, especially in gene expression. Generally, CNVs are more recent events than SNPs as they are still segregating within the population, showing 2.5 times faster evolution rate in the porcine genome (Paudel et al., 2015). Copy number variable genes in the porcine genome were suggested as candidates for selection related to traits such as coat color (Rubin et al., 2012; Xu et al., 2020), backfat thickness (Schiavo et al., 2014), fatty acid composition, growth (Revilla et al., 2017), and reproduction (Zheng et al., 2020). Therefore, comparing CNV can be an effective strategy for identifying recently accelerated differentiation between wild boar and domesticated pigs.

However, the number of individuals and populations in most of previous studies was not enough to suggest differentiated genomic regions between pig populations, such as indigenous breeds and wild boars. Furthermore, the credibility and resolution of CNVs were limited by using SNP chip or aligning on an older version of genome assembly. Here, I defined porcine CNVs from 328 individuals in 56 breeds, the largest population that represents their CNVs, including wild boar and domesticated and indigenous populations from broad area in Europe and Asia. I expected that my study on the comparison of pig CNVs between domesticated and wild would improve further understanding of the evolution of *Sus scrofa*.

3.3. Materials and Methods

3.3.1. Sample collection

The study population consisted of 328 individuals consist of 130 females and 198 males from 56 pig populations. The whole-genome sequencing (WGS) of wild boar and domesticated breeds were collected from Europe and Asia. 313 genomes were publicly available and sequenced using Illumina paired-end library and from SRA database (Table S1). 15 genomes including 5 Duroc, 5 Woori-Heukdon and 5 Korean Native were newly sequenced in this study. The 15 Blood samples were collected during routine veterinary treatments with the logistical support under the ethical approval of National Institute of Animal Science, Republic of Korea (NIAS20212224). All of the experimental protocols were approved by National Institute of Animal Science, Republic of Korea (NIAS20212224).

The 56 *Sus scrofa* populations were classified into five groups, European domesticated (ED), Asian domesticated (AD), European Wild Boar (EW), Asian Wild (AW), and Near Eastern Wild (NEW) as follows: i) 109 individuals of ED including 1 Angler Sattelschwein, 11 Berkshire, 1 British Saddleback, 1 Bunte Bentheimer, 2 Casertana, 1 Chato Murciano, 17 Duroc, 1 Gloucester Old Spot, 3 Hampshire, 4 Iberian, 4 Landrace, 37 Large White (Yorkshire), 2 Leping Spotted, 1 Linderodsvin, 5 Mangalica, 2 Middle White, 1 Nero Siciliano, 13 Pietrain and 2 Tamworth; ii) 120 individuals of AD including 6 Bamaxiang, 7 Bamei, 6 Baoshan, 3 Enshi black, 21 Erhualian, 6 Hetao, 3 Jiangquhai, 9 Jinhua, 5 Korean native, 5 Woori-Heukdon, 6 Laiwu, 6 Luchuan, 10 Meishan, 6 Min, 7 Neijiang, 6 Rongchang, 3 Tongcheng, 2 Wannan Spotted, 2 Xiang, and 1 Zang; iii) 20 individuals of EW including 12 Dutch, 1 French, 4 Italian, 2 Spanish and 1 Swiss wild boar; iv) 77 individuals of AW including 65 Chinese, 1 Japanese, 10 Korean, and 1 Russian wild boar; v) 2 Near Eastern wild boar. The additional information of samples is described in Table 3.1.

Table 3.1. Sample information and alignment statistics.

BioSample	Name	Population (Location)	Group	Sex	MappingRate	Coverage	Mean Depth	Number of Autosomal CNVs	Size of Autosomal DUP (bp)	Size of Autosomal DEL (bp)
SAMEA3497824	ANG1	Angler Sattelschwein	ED	F	0.99	0.96	11.12	1384	4.7.E+06	1.3.E+07
SAMN04440479	BAM1	Bamei	AD	F	0.99	0.97	55.96	3225	1.4.E+07	1.3.E+07
SAMN06348392	BAM2	Bamei	AD	M	0.99	0.97	19.38	2134	5.2.E+07	1.3.E+07
SAMN06348393	BAM3	Bamei	AD	M	0.99	0.97	19.63	1480	9.3.E+06	9.2.E+06
SAMN06348414	BAM4	Bamei	AD	M	0.99	0.97	24.09	1780	8.3.E+06	1.5.E+07
SAMN06348415	BAM5	Bamei	AD	M	0.99	0.97	20.01	1345	8.0.E+06	6.8.E+06
SAMN06348416	BAM6	Bamei	AD	M	0.97	0.97	21.98	1566	1.1.E+07	1.0.E+07
SAMN06348417	BAM7	Bamei	AD	M	0.99	0.97	20.65	1575	1.1.E+07	8.2.E+06
SAMN06349454	BAO1	Baoshan	AD	M	0.99	0.97	21.74	1584	7.8.E+06	9.5.E+06
SAMN06349455	BAO2	Baoshan	AD	M	0.99	0.97	21.33	1533	7.7.E+06	1.4.E+07
SAMN06349456	BAO3	Baoshan	AD	M	0.99	0.97	21.47	1497	8.0.E+06	8.8.E+06
SAMN06349457	BAO4	Baoshan	AD	F	0.99	0.97	21.51	1545	6.1.E+06	1.1.E+07
SAMN06349458	BAO5	Baoshan	AD	F	0.99	0.96	24.60	1816	9.8.E+06	2.1.E+07
SAMN06349459	BAO6	Baoshan	AD	M	0.99	0.96	23.37	1518	5.6.E+06	9.5.E+06
SAMEA3497827	BER1	Berkshire	ED	F	1.00	0.96	11.27	1142	4.0.E+06	7.3.E+06
SAMN03566761	BER10	Berkshire	ED	F	1.00	0.96	5.18	408	2.4.E+06	2.3.E+06
SAMN04440475	BER11	Berkshire	ED	F	0.99	0.98	73.99	2627	1.3.E+07	8.9.E+06

SAMEA3497828	BER2	Berkshire	ED	M	1.00	0.96	9.51	1212	5.1.E+06	7.5.E+06
SAMN03566754	BER3	Berkshire	ED	F	0.99	0.97	8.56	791	5.7.E+06	4.7.E+06
SAMN03566755	BER4	Berkshire	ED	F	1.00	0.97	9.22	902	4.4.E+06	5.8.E+06
SAMN03566756	BER5	Berkshire	ED	F	0.99	0.97	10.29	886	4.5.E+06	5.2.E+06
SAMN03566757	BER6	Berkshire	ED	F	1.00	0.97	13.86	1508	5.2.E+06	2.9.E+07
SAMN03566758	BER7	Berkshire	ED	F	1.00	0.97	10.19	809	4.3.E+06	4.0.E+06
SAMN03566759	BER8	Berkshire	ED	F	1.00	0.97	9.72	1805	7.9.E+06	2.1.E+07
SAMN03566760	BER9	Berkshire	ED	F	0.99	0.97	10.38	916	4.6.E+06	5.3.E+06
SAMN02298127	BMX1	Bamaxiang	AD	F	0.99	0.96	22.97	5401	1.1.E+07	5.9.E+07
SAMN02298128	BMX2	Bamaxiang	AD	F	0.99	0.96	24.16	5269	1.1.E+07	6.3.E+07
SAMN02298129	BMX3	Bamaxiang	AD	F	0.99	0.96	22.45	3572	1.1.E+07	1.8.E+07
SAMN02298130	BMX4	Bamaxiang	AD	M	0.99	0.96	23.04	5075	1.1.E+07	6.0.E+07
SAMN02298131	BMX5	Bamaxiang	AD	F	0.99	0.96	23.03	3802	1.1.E+07	2.1.E+07
SAMN02298132	BMX6	Bamaxiang	AD	F	0.99	0.96	23.00	4475	1.0.E+07	3.7.E+07
SAMEA3497830	BRI1	British Saddleback	ED	M	0.99	0.96	10.85	1064	5.1.E+06	6.4.E+06
SAMEA3497826	BUN1	Bunte Bentheimer	ED	M	0.99	0.97	13.45	1752	6.4.E+06	1.7.E+07
SAMEA3497832	CAS1	Casertana	ED	F	0.99	0.96	10.10	813	5.1.E+06	3.6.E+06
SAMEA3497835	CAS2	Casertana	ED	F	0.99	0.97	10.41	663	3.3.E+06	7.4.E+06
SAMEA3497836	CHM1	Chato Murciano	ED	M	0.99	0.96	8.06	713	3.9.E+06	3.1.E+06
SAMEA3497837	DUR1	Duroc	ED	M	0.99	0.97	12.33	1091	4.1.E+06	6.1.E+06

SAMEA3497838	DUR2	Duroc	ED	M	1.00	0.96	11.78	2433	3.5.E+06	3.8.E+07
SAMEA3497839	DUR3	Duroc	ED	M	1.00	0.95	5.91	592	3.9.E+06	3.4.E+06
SAMEA3497840	DUR4	Duroc	ED	M	1.00	0.95	7.44	1548	1.2.E+08	1.1.E+07
SAMN00005058	DUR5	Duroc	ED	M	0.97	0.96	7.23	193	7.7.E+05	1.1.E+06
SAMN03031126	DUR6	Duroc	ED	F	0.99	0.96	6.14	403	3.3.E+06	2.5.E+06
SAMN03031127	DUR7	Duroc	ED	F	0.99	0.97	12.91	782	4.6.E+06	3.1.E+06
SAMN03031128	DUR8	Duroc	ED	F	0.99	0.97	12.13	933	4.6.E+06	5.0.E+06
SAMN09930402	DUR9	Duroc	ED	M	0.99	0.97	20.47	1089	8.1.E+06	6.0.E+06
SAMN09930403	DUR10	Duroc	ED	M	0.99	0.97	30.66	2000	1.3.E+07	2.7.E+07
SAMN12122743	DUR11	Duroc	ED	M	0.99	0.97	29.45	1454	9.7.E+06	1.2.E+07
SAMN12122744	DUR12	Duroc	ED	M	0.99	0.97	30.66	1983	1.3.E+07	2.7.E+07
SAMN28745316	DUR13	Duroc	ED	M	0.99	0.97	46.61	1781	9.0.E+06	1.0.E+07
SAMN28745317	DUR14	Duroc	ED	M	0.99	0.97	51.12	1883	7.9.E+06	1.3.E+07
SAMN28745318	DUR15	Duroc	ED	M	0.99	0.98	33.16	1573	7.1.E+06	1.0.E+07
SAMN28745319	DUR16	Duroc	ED	M	0.99	0.98	28.99	1344	7.7.E+06	8.4.E+06
SAMN28745320	DUR17	Duroc	ED	M	0.99	0.98	45.71	1828	8.9.E+06	1.1.E+07
SAMN02298079	EHL1	Erhualian	AD	F	0.99	0.96	22.51	5218	1.1.E+07	5.3.E+07
SAMN02298080	EHL2	Erhualian	AD	F	0.99	0.96	23.17	3815	9.7.E+06	2.1.E+07
SAMN09930385	EHL3	Erhualian	AD	M	0.99	0.97	29.78	3610	8.2.E+06	4.7.E+07
SAMN09930386	EHL4	Erhualian	AD	M	0.99	0.97	30.27	3834	7.9.E+06	5.8.E+07
SAMN09930387	EHL5	Erhualian	AD	M	0.99	0.97	29.53	3613	7.7.E+06	4.3.E+07

SAMN09930388	EHL6	Erhualian	AD	F	0.98	0.97	29.87	2857	1.1.E+07	3.0.E+07
SAMN09930389	EHL7	Erhualian	AD	F	0.99	0.97	21.66	3705	9.1.E+06	4.3.E+07
SAMN09930390	EHL8	Erhualian	AD	F	0.99	0.97	22.05	3495	1.0.E+07	3.0.E+07
SAMN09930391	EHL9	Erhualian	AD	F	0.99	0.97	21.37	3982	9.3.E+06	3.9.E+07
SAMN09930392	EHL10	Erhualian	AD	F	0.99	0.97	21.51	3573	1.0.E+07	3.0.E+07
SAMN09930393	EHL11	Erhualian	AD	M	0.99	0.97	22.04	4298	9.6.E+06	6.0.E+07
SAMN09930394	EHL12	Erhualian	AD	M	0.99	0.97	21.73	4229	9.6.E+06	5.9.E+07
SAMN09930395	EHL13	Erhualian	AD	F	0.99	0.97	21.28	3972	1.1.E+07	4.3.E+07
SAMN09930396	EHL14	Erhualian	AD	F	0.99	0.97	20.29	4506	1.0.E+07	5.7.E+07
SAMN09930397	EHL15	Erhualian	AD	F	0.99	0.97	21.76	3683	9.6.E+06	3.3.E+07
SAMN09930398	EHL16	Erhualian	AD	F	0.99	0.97	21.36	3909	8.5.E+06	4.0.E+07
SAMN09930399	EHL17	Erhualian	AD	F	0.99	0.97	22.03	3796	9.8.E+06	4.3.E+07
SAMN09930400	EHL18	Erhualian	AD	F	0.99	0.97	21.46	3792	1.0.E+07	3.3.E+07
SAMN09930401	EHL19	Erhualian	AD	F	0.99	0.97	22.12	3717	9.2.E+06	2.8.E+07
SAMN12122745	EHL20	Erhualian	AD	F	0.99	0.96	22.42	5267	1.1.E+07	5.4.E+07
SAMN12122746	EHL21	Erhualian	AD	F	0.99	0.96	23.11	3774	9.4.E+06	2.0.E+07
SAMN04538376	ENB1	Enshi black	AD	M	0.99	0.96	15.67	3331	2.5.E+08	1.7.E+07
SAMN04538598	ENB2	Enshi black	AD	M	0.99	0.96	14.73	3175	1.7.E+08	2.7.E+07
SAMN04538599	ENB3	Enshi black	AD	M	0.99	0.96	12.65	4136	2.3.E+08	2.2.E+07
SAMEA3497842	GLO1	Gloucester Old Spot	ED	M	0.99	0.95	8.23	5691	2.0.E+08	5.7.E+07
SAMEA3497843	HAM1	Hampshire	ED	M	1.00	0.96	9.00	846	8.8.E+06	2.9.E+06

SAMEA3497844	HAM2	Hampshire	ED	M	0.99	0.96	8.38	2018	4.8.E+07	1.0.E+07
SAMN04440474	HAM3	Hampshire	ED	F	0.99	0.98	66.71	2438	1.0.E+07	9.0.E+06
SAMN02298115	HT1	Hetao	AD	F	0.99	0.96	21.74	4631	8.2.E+06	3.1.E+07
SAMN02298116	HT2	Hetao	AD	M	1.00	0.96	21.18	3893	7.4.E+06	2.2.E+07
SAMN02298117	HT3	Hetao	AD	M	0.99	0.97	17.83	4607	7.3.E+06	3.7.E+07
SAMN02298118	HT4	Hetao	AD	M	0.99	0.97	20.99	3878	8.5.E+06	2.4.E+07
SAMN02298119	HT5	Hetao	AD	F	0.99	0.96	21.58	5223	8.2.E+06	4.8.E+07
SAMN02298120	HT6	Hetao	AD	F	1.00	0.96	20.93	6328	7.2.E+06	5.6.E+07
SAMN02904857	IBE1	Iberian	ED	M	0.99	0.95	9.93	2137	4.0.E+06	3.1.E+07
SAMN03421607	IBE2	Iberian	ED	M	0.97	0.97	13.32	774	3.2.E+06	4.2.E+06
SAMN05362554	IBE3	Iberian	ED	M	0.99	0.97	11.48	716	3.0.E+06	4.0.E+06
SAMN06895012	IBE4	Iberian	ED	M	0.96	0.97	11.53	672	3.3.E+06	3.4.E+06
SAMN06349462	JIN1	Jinhua	AD	M	0.99	0.96	20.75	1578	6.0.E+06	9.4.E+06
SAMN06349463	JIN2	Jinhua	AD	F	0.99	0.96	24.93	1977	7.3.E+06	1.8.E+07
SAMN06349464	JIN3	Jinhua	AD	M	0.99	0.97	21.47	1605	6.8.E+06	1.5.E+07
SAMN06349465	JIN4	Jinhua	AD	M	0.99	0.97	22.85	1567	7.7.E+06	9.3.E+06
SAMN06349466	JIN5	Jinhua	AD	M	0.99	0.96	20.71	1813	6.0.E+06	1.7.E+07
SAMN06349467	JIN6	Jinhua	AD	M	0.99	0.97	22.09	1901	1.0.E+07	1.5.E+07
SAMEA3497793	JIN7	Jinhua	AD	M	0.99	0.96	8.21	1322	3.1.E+07	6.6.E+06
SAMEA3497794	JIN8	Jinhua	AD	F	0.99	0.96	9.08	1283	4.1.E+06	2.1.E+07
SAMN04440480	JIN9	Jinhua	AD	M	0.99	0.97	69.29	588	3.1.E+06	2.8.E+06

SAMEA3497795	JQH1	Jiangquhai	AD	F	0.99	0.96	10.68	3734	8.4.E+06	4.8.E+07
SAMEA3497796	JQH2	Jiangquhai	AD	M	0.99	0.96	7.31	736	3.5.E+06	2.8.E+06
SAMEA3497797	JQH3	Jiangquhai	AD	M	0.99	0.96	7.72	744	4.1.E+06	4.1.E+06
SAMN28745321	KNP1	Korean Native	AD	M	0.99	0.97	30.41	2032	1.1.E+07	1.2.E+07
SAMN28745322	KNP2	Korean Native	AD	M	0.99	0.97	34.82	2370	1.0.E+07	1.5.E+07
SAMN28745323	KNP3	Korean Native	AD	M	0.99	0.98	43.61	2978	9.3.E+06	3.6.E+07
SAMN28745324	KNP4	Korean Native	AD	M	0.99	0.98	48.99	2400	1.2.E+07	1.4.E+07
SAMN28745325	KNP5	Korean Native	AD	M	0.99	0.97	29.36	2111	9.3.E+06	1.2.E+07
SAMN28745312	KWH1	Woori-Heukdon	AD	M	0.99	0.98	28.30	1992	9.5.E+06	1.3.E+07
SAMN28745313	KWH2	Woori-Heukdon	AD	M	0.99	0.98	42.77	1434	8.2.E+06	9.7.E+06
SAMN28745314	KWH3	Woori-Heukdon	AD	M	0.99	0.98	30.15	1996	1.1.E+07	1.4.E+07
SAMN28745315	KWH4	Woori-Heukdon	AD	M	0.99	0.98	44.98	1673	7.8.E+06	1.2.E+07
SAMN28745297	KWH5	Woori-Heukdon	AD	M	1.00	0.98	26.97	1740	7.1.E+06	1.8.E+07
SAMEA3497847	LAN1	Landrace	ED	M	1.00	0.96	9.37	1911	1.2.E+08	3.7.E+07
SAMEA3497850	LAN2	Landrace	ED	F	1.00	0.96	7.49	1131	3.6.E+06	1.2.E+07
SAMEA3497851	LAN3	Landrace	ED	M	0.98	0.97	9.55	641	4.0.E+06	3.6.E+06
SAMN04440476	LAN4	Landrace	ED	F	0.99	0.98	58.75	2372	1.2.E+07	9.7.E+06
SAMEA3497798	LEP1	Leping Spotted	ED	F	0.99	0.96	9.85	1327	5.7.E+06	8.8.E+06
SAMEA3497799	LEP2	Leping Spotted	ED	F	0.99	0.96	12.50	1299	7.2.E+06	2.2.E+07

SAMEA3497852	LIN1	Linderodsvin	ED	F	1.00	0.96	11.10	1863	3.5.E+06	1.4.E+07
SAMN02298087	LUC1	Luchuan	AD	F	1.00	0.96	19.50	8453	3.3.E+07	1.7.E+08
SAMN02298088	LUC2	Luchuan	AD	F	1.00	0.96	23.27	7688	1.2.E+07	1.5.E+08
SAMN02298089	LUC3	Luchuan	AD	F	0.99	0.96	22.91	6667	1.4.E+07	1.3.E+08
SAMN02298090	LUC4	Luchuan	AD	F	0.99	0.96	22.56	4288	1.3.E+07	5.5.E+07
SAMN02298091	LUC5	Luchuan	AD	F	0.99	0.97	23.39	8114	4.5.E+07	1.4.E+08
SAMN02298092	LUC6	Luchuan	AD	F	0.99	0.96	20.64	9169	2.0.E+08	1.6.E+08
SAMN02298133	LWU1	Laiwu	AD	M	0.99	0.97	22.14	3189	8.2.E+06	1.8.E+07
SAMN02298134	LWU2	Laiwu	AD	M	0.99	0.97	20.92	3046	9.0.E+06	1.7.E+07
SAMN02298135	LWU3	Laiwu	AD	M	0.99	0.97	22.73	3624	1.1.E+07	3.2.E+07
SAMN02298136	LWU4	Laiwu	AD	M	0.99	0.97	21.79	3400	9.2.E+06	3.0.E+07
SAMN02298137	LWU5	Laiwu	AD	M	0.99	0.97	22.13	6063	1.9.E+08	1.2.E+08
SAMN02298138	LWU6	Laiwu	AD	M	0.99	0.97	21.98	4201	1.6.E+07	6.3.E+07
SAMEA3497854	MAN1	Mangalica	ED	F	0.99	0.96	8.67	674	2.6.E+06	2.7.E+06
SAMEA3497855	MAN2	Mangalica	ED	M	0.99	0.97	9.81	711	3.7.E+06	3.7.E+06
SAMN02665304	MAN3	Mangalica	ED	M	0.98	0.97	15.91	438	3.4.E+06	5.8.E+06
SAMN02665305	MAN4	Mangalica	ED	M	0.99	0.97	12.22	1088	3.8.E+06	1.3.E+07
SAMN02665306	MAN5	Mangalica	ED	M	0.99	0.96	11.59	1003	3.8.E+06	1.2.E+07
SAMEA3497800	MEI1	Meishan	AD	M	0.99	0.97	9.90	1952	1.7.E+08	3.9.E+07
SAMN04440481	MEI10	Meishan	AD	F	0.99	0.97	70.31	921	4.8.E+06	3.5.E+06
SAMEA3497801	MEI2	Meishan	AD	M	0.99	0.97	10.27	1247	6.3.E+06	6.2.E+06

SAMEA3497802	MEI3	Meishan	AD	F	0.99	0.96	9.07	1512	5.2.E+06	9.1.E+06
SAMEA3497803	MEI4	Meishan	AD	M	0.99	0.96	8.89	2459	6.2.E+06	2.0.E+07
SAMEA3497804	MEI5	Meishan	AD	M	0.99	0.96	7.46	846	3.3.E+06	7.9.E+06
SAMEA3497805	MEI6	Meishan	AD	M	0.99	0.96	8.38	1158	6.0.E+06	5.2.E+06
SAMEA3497806	MEI7	Meishan	AD	M	0.99	0.97	10.41	1150	6.0.E+06	5.5.E+06
SAMEA3497807	MEI8	Meishan	AD	M	0.99	0.96	8.69	1052	2.0.E+07	2.1.E+06
SAMEA3497808	MEI9	Meishan	AD	M	0.99	0.96	8.71	1142	5.6.E+06	6.0.E+06
SAMEA3497809	MID1	Middle White	ED	M	0.99	0.96	13.95	5396	1.3.E+07	9.8.E+07
SAMEA3497856	MID2	Middle White	ED	F	0.99	0.96	11.13	1168	5.7.E+06	6.0.E+06
SAMN02298121	MIN1	Min	AD	F	1.00	0.96	21.74	5368	8.7.E+06	6.6.E+07
SAMN02298122	MIN2	Min	AD	M	0.99	0.97	21.81	4321	8.6.E+06	3.4.E+07
SAMN02298123	MIN3	Min	AD	M	0.99	0.96	21.95	5005	1.3.E+07	8.8.E+07
SAMN02298124	MIN4	Min	AD	M	1.00	0.96	20.40	3695	7.5.E+06	2.3.E+07
SAMN02298125	MIN5	Min	AD	M	0.99	0.97	23.18	4054	1.0.E+07	3.1.E+07
SAMN02298126	MIN6	Min	AD	M	0.99	0.97	22.21	2713	9.2.E+06	2.5.E+07
SAMN01894448	NEI1	Neijiang	AD	F	1.00	0.92	5.43	1390	1.5.E+08	6.4.E+06
SAMN06393132	NEI2	Neijiang	AD	M	0.99	0.97	22.73	1935	9.1.E+06	1.3.E+07
SAMN06393133	NEI3	Neijiang	AD	M	0.99	0.97	22.88	1649	9.9.E+06	7.8.E+06
SAMN06393134	NEI4	Neijiang	AD	M	0.99	0.97	23.31	1718	6.7.E+06	1.1.E+07
SAMN06393485	NEI5	Neijiang	AD	M	0.99	0.97	20.64	1658	8.2.E+06	9.5.E+06
SAMN06394064	NEI6	Neijiang	AD	M	0.99	0.97	20.92	1773	9.4.E+06	1.0.E+07

SAMN06394627	NEI7	Neijiang	AD	M	0.99	0.97	22.23	2214	9.1.E+06	1.6.E+07
SAMN08035066	NES1	Nero Siciliano	ED	M	0.99	0.97	31.15	1692	8.0.E+06	1.0.E+07
SAMEA3376934	PIE1	Pietrain	ED	F	0.99	0.96	6.48	401	2.6.E+06	3.5.E+06
SAMEA3376936	PIE2	Pietrain	ED	F	0.99	0.96	10.62	1108	5.1.E+06	6.9.E+06
SAMEA3376937	PIE3	Pietrain	ED	F	1.00	0.95	11.32	3147	4.9.E+06	4.0.E+07
SAMEA3376938	PIE4	Pietrain	ED	F	1.00	0.95	11.85	3529	4.4.E+06	3.7.E+07
SAMEA3376939	PIE5	Pietrain	ED	F	0.99	0.96	12.95	1438	5.7.E+06	8.8.E+06
SAMEA3376940	PIE6	Pietrain	ED	M	0.99	0.97	8.97	6459	1.3.E+08	3.3.E+06
SAMEA3376941	PIE7	Pietrain	ED	M	0.99	0.95	9.19	1251	1.1.E+08	1.1.E+06
SAMEA3376942	PIE8	Pietrain	ED	M	0.99	0.97	9.36	1051	6.2.E+06	8.1.E+06
SAMEA3376943	PIE9	Pietrain	ED	M	0.99	0.97	9.48	891	4.7.E+06	8.5.E+06
SAMEA3376944	PIE10	Pietrain	ED	M	0.99	0.96	9.26	21441	3.3.E+08	1.2.E+08
SAMEA3497791	PIE11	Pietrain	ED	M	1.00	0.95	7.15	2070	1.1.E+08	1.3.E+07
SAMEA3497860	PIE12	Pietrain	ED	M	0.99	0.93	6.42	82	1.1.E+06	1.4.E+05
SAMN04440477	PIE13	Pietrain	ED	F	0.99	0.98	56.74	2244	1.1.E+07	7.7.E+06
SAMN02460623	RON1	Rongchang	AD	M	0.99	0.96	7.31	433	2.8.E+06	1.5.E+06
SAMN02460625	RON2	Rongchang	AD	F	0.99	0.95	6.86	513	4.7.E+06	1.6.E+06
SAMN02460626	RON3	Rongchang	AD	M	0.99	0.96	7.69	587	5.4.E+06	2.4.E+06
SAMN02460627	RON4	Rongchang	AD	M	0.99	0.95	6.21	502	2.4.E+06	3.1.E+06
SAMN03331745	RON5	Rongchang	AD	M	0.99	0.95	5.82	305	2.3.E+06	1.3.E+06
SAMN04440482	RON6	Rongchang	AD	M	0.99	0.97	60.90	735	3.6.E+06	3.2.E+06

SAMEA3497862	TAM1	Tamworth	ED	F	0.99	0.96	10.29	659	3.1.E+06	3.6.E+06
SAMEA3497863	TAM2	Tamworth	ED	M	0.99	0.97	11.78	798	3.9.E+06	6.3.E+06
SAMN02646543	TON1	Tongcheng	AD	M	1.00	0.94	6.94	2272	7.9.E+07	1.7.E+07
SAMN02646544	TON2	Tongcheng	AD	F	0.99	0.94	5.59	2150	1.6.E+08	1.4.E+07
SAMN02646545	TON3	Tongcheng	AD	M	0.99	0.95	7.06	1690	6.4.E+07	6.9.E+06
SAMEA3497810	WAN1	Wannan Spotted	AD	F	0.99	0.96	8.97	1081	4.9.E+06	6.0.E+06
SAMEA3497811	WAN2	Wannan Spotted	AD	F	0.99	0.96	8.45	1029	4.5.E+06	6.0.E+06
SAMEA3497864	WDU1	Dutch Wild	EW	F	0.98	0.96	9.46	685	2.1.E+06	4.1.E+06
SAMEA3497865	WDU2	Dutch Wild	EW	F	0.99	0.96	11.30	1360	2.2.E+06	2.6.E+07
SAMEA3497866	WDU3	Dutch Wild	EW	M	0.99	0.96	9.17	1289	2.7.E+06	2.6.E+07
SAMEA3497867	WDU4	Dutch Wild	EW	M	1.00	0.96	11.38	2223	3.7.E+06	1.8.E+07
SAMEA3497868	WDU5	Dutch Wild	EW	M	0.99	0.96	9.97	846	5.0.E+06	5.7.E+06
SAMEA3497869	WDU6	Dutch Wild	EW	M	0.98	0.97	16.36	631	3.0.E+06	4.1.E+06
SAMEA3497870	WDU7	Dutch Wild	EW	M	1.00	0.94	6.26	70	1.3.E+06	7.2.E+05
SAMEA3497871	WDU8	Dutch Wild	EW	F	1.00	0.96	7.76	1633	2.2.E+06	3.2.E+07
SAMEA3497872	WDU9	Dutch Wild	EW	F	0.99	0.96	8.31	967	3.0.E+06	9.3.E+06
SAMEA3497873	WDU10	Dutch Wild	EW	M	0.99	0.96	7.94	577	1.5.E+06	3.4.E+06
SAMEA3497874	WDU11	Dutch Wild	EW	M	0.99	0.97	12.14	1175	2.1.E+06	1.5.E+07
SAMEA3497875	WDU12	Dutch Wild	EW	M	0.99	0.97	10.63	850	4.3.E+06	3.7.E+06
SAMEA3497876	WFR1	French Wild	EW	M	0.99	0.96	9.41	1234	1.0.E+07	7.1.E+06

SAMEA3497879	WIT1	Italian Wild	EW	M	1.00	0.96	10.75	1789	5.8.E+06	1.4.E+07
SAMEA3497886	WIT2	Italian Wild	EW	F	0.99	0.97	12.27	902	4.4.E+06	6.9.E+06
SAMEA3497887	WIT3	Italian Wild	EW	M	0.98	0.97	11.91	1112	4.5.E+06	1.8.E+07
SAMEA3497888	WIT4	Italian Wild	EW	M	0.99	0.96	10.69	1895	1.8.E+08	7.4.E+07
SAMEA3497823	WJP1	Japanese Wild	AW	F	1.00	0.95	11.14	4809	7.2.E+07	5.4.E+07
SAMN03031171	WKO1	Korean Wild	AW	M	1.00	0.96	11.34	4035	1.3.E+08	3.7.E+07
SAMN03031172	WKO2	Korean Wild	AW	M	0.99	0.96	7.96	787	1.9.E+07	2.5.E+06
SAMN03031173	WKO3	Korean Wild	AW	M	1.00	0.95	10.83	1229	2.0.E+07	3.3.E+06
SAMN03031174	WKO4	Korean Wild	AW	F	1.00	0.96	11.55	1338	1.6.E+06	1.1.E+07
SAMN03031175	WKO5	Korean Wild	AW	M	1.00	0.95	10.49	2189	7.6.E+06	2.2.E+07
SAMN03031176	WKO6	Korean Wild	AW	F	1.00	0.96	9.86	1159	1.8.E+06	1.1.E+07
SAMN03031177	WKO7	Korean Wild	AW	M	0.99	0.96	7.59	431	5.9.E+06	1.3.E+06
SAMN03031178	WKO8	Korean Wild	AW	M	0.99	0.96	10.44	2251	1.2.E+07	2.0.E+07
SAMN03031179	WKO9	Korean Wild	AW	M	0.99	0.96	10.42	472	2.1.E+06	2.6.E+06
SAMN03031180	WKO10	Korean Wild	AW	M	1.00	0.96	9.42	2177	9.9.E+06	1.3.E+07
SAMEA3497821	WNC1	Northern Chinese Wild	AW	M	0.99	0.96	10.10	3078	1.0.E+07	3.0.E+07
SAMEA3497822	WNC2	Northern Chinese Wild	AW	M	0.97	0.97	12.10	761	4.9.E+06	3.2.E+06
SAMEA3497884	WNE1	Near Eastern Wild	NEW	M	0.99	0.97	11.02	977	3.2.E+06	7.5.E+06
SAMEA3497885	WNE2	Near Eastern Wild	NEW	F	0.99	0.97	9.97	777	3.9.E+06	3.4.E+06
SAMN05362551	WRU1	Russian-	AW	M	0.97	0.94	5.90	2988	5.5.E+07	1.0.E+07

		Primorskiy Kray								
SAMEA3497815	WSC1	Southern Chinese Wild	AW	M	0.99	0.92	6.03	97	3.0.E+06	1.3.E+05
SAMEA3497816	WSC2	Southern Chinese Wild	AW	F	0.99	0.96	10.29	1931	7.5.E+06	1.3.E+07
SAMEA3497818	WSC3	Southern Chinese Wild	AW	M	0.99	0.97	30.04	2479	1.0.E+07	1.1.E+07
SAMEA3497819	WSC4	Southern Chinese Wild	AW	M	0.99	0.97	12.43	1368	6.3.E+06	6.7.E+06
SAMN01894459	WSC5	Southern Chinese Wild	AW	F	0.99	0.94	5.07	1087	3.1.E+06	1.6.E+07
SAMN02298081	WSC6	Southern Chinese Wild	AW	M	0.99	0.97	22.90	3681	1.1.E+07	2.0.E+07
SAMN02298082	WSC7	Southern Chinese Wild	AW	M	0.99	0.97	23.82	3834	1.2.E+07	2.5.E+07
SAMN02298083	WSC8	Southern Chinese Wild	AW	F	0.99	0.96	22.68	5624	1.3.E+07	7.3.E+07
SAMN02298084	WSC9	Southern Chinese Wild	AW	M	1.00	0.97	17.48	6225	2.9.E+07	9.1.E+07
SAMN02298085	WSC10	Southern Chinese Wild	AW	M	0.99	0.96	16.92	9612	1.7.E+08	1.6.E+08
SAMN02298086	WSC11	Southern Chinese Wild	AW	M	0.99	0.96	17.96	9725	2.1.E+08	1.7.E+08
SAMN02904855	WSP1	Spainish Wild	EW	M	0.98	0.97	11.58	731	4.2.E+06	3.8.E+06
SAMN05362552	WSP2	Spainish Wild	EW	M	0.99	0.97	11.81	469	2.5.E+06	2.4.E+06
SAMEA3497877	WSW1	Swiss Wild	EW	M	0.99	0.96	8.41	619	2.6.E+06	5.2.E+06
SAMN02298093	WT1	Tibetan Wild	AW	F	0.99	0.96	22.76	4447	8.3.E+06	4.3.E+07
SAMN02298094	WT2	Tibetan Wild	AW	M	1.00	0.97	21.48	4430	9.8.E+06	2.9.E+07

SAMN02298095	WT3	Tibetan Wild	AW	M	0.99	0.97	18.78	5695	2.5.E+07	8.6.E+07
SAMN02298096	WT4	Tibetan Wild	AW	M	0.99	0.97	17.70	4162	6.9.E+06	4.8.E+07
SAMN02298097	WT5	Tibetan Wild	AW	F	0.99	0.96	18.70	5122	7.7.E+06	6.6.E+07
SAMN02298098	WT6	Tibetan Wild	AW	M	0.99	0.97	23.96	3224	9.7.E+06	2.4.E+07
SAMN12122795	WT7	Tibetan Wild	AW	M	0.99	0.97	20.45	1302	7.1.E+06	9.5.E+06
SAMN12122796	WT8	Tibetan Wild	AW	M	0.99	0.97	24.86	1665	9.6.E+06	1.0.E+07
SAMN12122797	WT9	Tibetan Wild	AW	M	0.99	0.97	22.83	1705	1.2.E+07	8.9.E+06
SAMN12122798	WT10	Tibetan Wild	AW	M	0.99	0.97	22.19	1469	9.1.E+06	8.0.E+06
SAMN12122799	WT11	Tibetan Wild	AW	M	0.99	0.97	26.51	1820	8.3.E+06	1.7.E+07
SAMN12122800	WT12	Tibetan Wild	AW	F	0.99	0.97	22.50	2428	5.6.E+06	3.6.E+07
SAMN01894407	WTGN1	Gannan Tibetan Wild	AW	F	1.00	0.93	5.09	1691	1.1.E+08	4.8.E+07
SAMN02298111	WTGS1	Gansu Tibetan Wild	AW	M	1.00	0.96	22.15	6434	8.8.E+06	6.6.E+07
SAMN02298112	WTGS2	Gansu Tibetan Wild	AW	F	0.99	0.96	19.74	6020	8.0.E+06	6.0.E+07
SAMN02298113	WTGS3	Gansu Tibetan Wild	AW	M	0.99	0.96	20.40	5425	9.8.E+06	6.5.E+07
SAMN02298114	WTGS4	Gansu Tibetan Wild	AW	M	0.99	0.96	20.70	5795	8.0.E+06	4.9.E+07
SAMN12122783	WTGS5	Gansu Tibetan Wild	AW	M	0.99	0.97	24.32	1894	1.0.E+07	1.2.E+07
SAMN12122784	WTGS6	Gansu Tibetan Wild	AW	M	0.99	0.97	25.04	1791	7.6.E+06	1.4.E+07
SAMN12122785	WTGS7	Gansu Tibetan Wild	AW	M	0.99	0.97	21.98	1955	9.6.E+06	2.3.E+07

SAMN12122786	WTGS8	Gansu Tibetan Wild	AW	F	0.99	0.96	21.16	1479	7.3.E+06	1.0.E+07
SAMN12122787	WTGS9	Gansu Tibetan Wild	AW	M	0.99	0.97	22.83	1377	7.1.E+06	8.2.E+06
SAMN12122788	WTGS10	Gansu Tibetan Wild	AW	M	0.99	0.97	22.66	1946	5.9.E+06	2.2.E+07
SAMN01894388	WTN1	Nyingchi Tibetan Wild	AW	M	1.00	0.95	6.41	898	3.5.E+06	1.2.E+07
SAMN01894391	WTN2	Nyingchi Tibetan Wild	AW	M	1.00	0.94	5.85	1489	7.9.E+07	4.9.E+07
SAMN01894434	WTSC1	Sichuan Tibetan Wild	AW	M	1.00	0.94	6.14	1042	1.5.E+07	7.3.E+06
SAMN01894436	WTSC2	Sichuan Tibetan Wild	AW	F	1.00	0.93	5.66	1353	8.5.E+07	1.2.E+07
SAMN02298105	WTSC3	Sichuan Tibetan Wild	AW	F	1.00	0.96	22.82	5200	8.7.E+06	4.3.E+07
SAMN02298106	WTSC4	Sichuan Tibetan Wild	AW	F	0.99	0.96	22.18	4565	8.7.E+06	3.1.E+07
SAMN02298107	WTSC5	Sichuan Tibetan Wild	AW	M	0.99	0.97	22.00	4366	9.8.E+06	3.7.E+07
SAMN02298108	WTSC6	Sichuan Tibetan Wild	AW	M	0.99	0.97	17.37	3071	9.3.E+06	2.0.E+07
SAMN02298109	WTSC7	Sichuan Tibetan Wild	AW	M	0.99	0.97	23.85	3735	1.3.E+07	2.2.E+07
SAMN02298110	WTSC8	Sichuan Tibetan Wild	AW	M	1.00	0.97	22.81	4717	1.0.E+07	3.5.E+07
SAMN12122789	WTSC9	Sichuan Tibetan Wild	AW	M	0.99	0.97	21.87	1652	6.9.E+06	1.5.E+07
SAMN12122790	WTSC10	Sichuan Tibetan Wild	AW	M	0.99	0.97	24.59	1705	6.9.E+06	1.5.E+07

SAMN12122791	WTSC11	Sichuan Tibetan Wild	AW	M	0.99	0.97	21.60	1738	9.0.E+06	9.1.E+06
SAMN12122792	WTSC12	Sichuan Tibetan Wild	AW	M	0.99	0.97	23.63	1721	1.1.E+07	1.2.E+07
SAMN12122793	WTSC13	Sichuan Tibetan Wild	AW	M	0.99	0.97	20.45	1302	7.1.E+06	9.5.E+06
SAMN12122794	WTSC14	Sichuan Tibetan Wild	AW	M	0.99	0.97	22.10	2259	1.0.E+07	2.1.E+07
SAMN01894367	WTY1	Yunnan Tibetan Wild	AW	F	1.00	0.94	5.27	1180	6.7.E+07	1.4.E+07
SAMN01894370	WTY2	Yunnan Tibetan Wild	AW	F	1.00	0.95	5.97	1424	1.4.E+08	3.4.E+07
SAMN02298099	WTY3	Yunnan Tibetan Wild	AW	M	0.99	0.97	23.81	4588	1.1.E+07	5.2.E+07
SAMN02298100	WTY4	Yunnan Tibetan Wild	AW	M	0.99	0.97	17.84	3814	1.6.E+07	5.0.E+07
SAMN02298101	WTY5	Yunnan Tibetan Wild	AW	M	0.99	0.97	23.57	3222	9.2.E+06	2.0.E+07
SAMN02298103	WTY6	Yunnan Tibetan Wild	AW	M	0.99	0.97	22.25	2994	1.0.E+07	1.7.E+07
SAMN02298104	WTY7	Yunnan Tibetan Wild	AW	M	0.99	0.97	22.64	3591	1.7.E+07	4.0.E+07
SAMN12122801	WTY8	Yunnan Tibetan Wild	AW	M	0.99	0.97	21.67	1569	8.4.E+06	9.4.E+06
SAMN12122802	WTY9	Yunnan Tibetan Wild	AW	M	0.99	0.97	24.79	1666	7.3.E+06	1.4.E+07
SAMN12122803	WTY10	Yunnan Tibetan Wild	AW	M	1.00	0.97	22.52	1711	6.8.E+06	1.6.E+07
SAMN12122804	WTY11	Yunnan Tibetan Wild	AW	M	0.99	0.97	21.56	1790	8.6.E+06	1.8.E+07

SAMN12122805	WTY12	Yunnan Tibetan Wild	AW	M	1.00	0.97	21.34	2279	9.0.E+06	1.9.E+07
SAMN12122806	WTY13	Yunnan Tibetan Wild	AW	M	0.99	0.97	21.19	1690	9.7.E+06	9.4.E+06
SAMEA3497812	XIA1	Xiang	AD	F	0.99	0.96	8.16	2252	7.5.E+06	1.7.E+07
SAMEA3497813	XIA2	Xiang	AD	M	0.99	0.96	7.97	2103	6.3.E+06	1.3.E+07
SAMEA3497853	YOR1	Large White	ED	M	1.00	0.96	9.68	10213	1.1.E+08	6.0.E+07
SAMN04440478	YOR2	Large White	ED	M	0.99	0.98	60.89	2504	1.0.E+07	9.7.E+06
SAMN12122747	YOR3	Large White	ED	F	0.99	0.97	18.62	2232	8.7.E+06	1.6.E+07
SAMN12122749	YOR4	Large White	ED	F	0.99	0.97	19.24	2300	9.1.E+06	2.8.E+07
SAMN12122750	YOR5	Large White	ED	M	0.99	0.97	19.26	2470	1.2.E+07	1.7.E+07
SAMN12122751	YOR6	Large White	ED	F	0.99	0.97	18.28	2357	9.1.E+06	2.3.E+07
SAMN12122752	YOR7	Large White	ED	F	0.99	0.97	19.74	2415	9.0.E+06	1.8.E+07
SAMN12122753	YOR8	Large White	ED	F	0.99	0.97	18.93	2407	8.0.E+06	2.5.E+07
SAMN12122754	YOR9	Large White	ED	F	0.99	0.96	16.92	2986	1.4.E+07	3.8.E+07
SAMN12122755	YOR10	Large White	ED	F	0.99	0.97	17.97	2750	1.1.E+07	4.3.E+07
SAMN12122756	YOR11	Large White	ED	M	0.99	0.97	19.21	2214	8.5.E+06	1.7.E+07
SAMN12122757	YOR12	Large White	ED	F	0.99	0.97	18.09	2715	1.1.E+07	4.4.E+07
SAMN12122758	YOR13	Large White	ED	F	0.99	0.97	20.35	1923	7.1.E+06	1.5.E+07
SAMN12122759	YOR14	Large White	ED	F	0.99	0.97	18.61	1972	1.1.E+07	2.6.E+07
SAMN12122760	YOR15	Large White	ED	F	0.99	0.97	18.03	2784	8.3.E+06	4.8.E+07
SAMN12122761	YOR16	Large White	ED	F	0.99	0.97	18.78	2276	1.0.E+07	1.6.E+07

SAMN12122762	YOR17	Large White	ED	F	0.99	0.97	19.33	2411	1.1.E+07	1.6.E+07
SAMN12122763	YOR18	Large White	ED	F	0.99	0.96	14.70	2303	1.4.E+07	3.0.E+07
SAMN12122764	YOR19	Large White	ED	F	0.99	0.97	20.23	2456	7.5.E+06	3.2.E+07
SAMN12122765	YOR20	Large White	ED	F	0.99	0.97	18.72	2223	9.4.E+06	1.7.E+07
SAMN12122766	YOR21	Large White	ED	F	0.99	0.97	18.71	2088	8.3.E+06	1.4.E+07
SAMN12122767	YOR22	Large White	ED	F	0.99	0.97	19.10	2264	8.8.E+06	2.5.E+07
SAMN12122768	YOR23	Large White	ED	F	0.99	0.97	18.00	2611	1.1.E+07	3.9.E+07
SAMN12122769	YOR24	Large White	ED	F	0.99	0.97	19.21	1826	8.7.E+06	1.4.E+07
SAMN12122770	YOR25	Large White	ED	F	0.99	0.97	18.01	2268	1.1.E+07	3.1.E+07
SAMN12122771	YOR26	Large White	ED	F	0.99	0.97	17.42	1677	8.1.E+06	1.7.E+07
SAMN12122772	YOR27	Large White	ED	F	0.99	0.97	24.24	3009	1.0.E+07	2.6.E+07
SAMN12122773	YOR28	Large White	ED	F	0.99	0.97	20.90	2406	9.7.E+06	1.8.E+07
SAMN12122774	YOR29	Large White	ED	F	0.99	0.97	18.59	2295	1.2.E+07	2.0.E+07
SAMN12122775	YOR30	Large White	ED	F	0.99	0.97	20.37	2352	1.1.E+07	1.8.E+07
SAMN12122776	YOR31	Large White	ED	F	0.99	0.96	16.27	7052	1.2.E+08	4.3.E+07
SAMN12122777	YOR32	Large White	ED	M	0.99	0.97	17.55	2809	1.4.E+07	3.9.E+07
SAMN12122778	YOR33	Large White	ED	F	0.99	0.96	16.46	3521	1.4.E+07	9.3.E+07
SAMN12122779	YOR34	Large White	ED	F	0.99	0.97	22.79	2742	7.8.E+06	2.4.E+07
SAMN12122780	YOR35	Large White	ED	F	0.99	0.97	20.42	2150	7.7.E+06	2.0.E+07
SAMN12122781	YOR36	Large White	ED	M	0.99	0.97	18.63	2570	1.0.E+07	3.9.E+07
SAMN12122782	YOR37	Large White	ED	F	0.99	0.97	20.68	2392	8.2.E+06	2.5.E+07

SAMEA3497814	ZAN1	Zang	AD	M	0.99	0.97	9.03	973	6.6.E+06	5.0.E+06

3.3.2. Whole genome sequencing

Fifteen genomes including 5 Duroc, 5 Woori-Heukdon and 5 Korean Native were newly sequenced in this study. Blood samples were collected for DNA extraction by Wizard® Genomic DNA Purification Kit (Promega) from National Institute of Animal Science, Rural National Institute of Animal Science, Republic of Korea. Library construction was performed for each individual using 2µg of genomic DNA with Illumina TruSeq PCR-free (550) Kit. Sequencing was performed to generate 2 x 151 paired-end reads on the Illumina NovaSeq 6000 platform.

3.3.3. Whole genome sequence alignment

After quality control checking of raw reads using FastQC-0.11.8 (Andrews, 2017), adapter and low-quality bases of reads were trimmed by Trimmomatic-0.39 (Bolger et al., 2014). After checking the trimming results and quality of trimmed reads, the trimmed reads were mapped using BWA-0.7.17 MEM (Li & Durbin, 2009) to reference genome Sscrofa11.1 assembly (Warr et al., 2020). The outputs of the sequence alignment map (SAM) were sorted, indexed, and compressed to binary format (BAM) by Samtools-1.9 (Li et al., 2009). The duplicates in BAM files were marked using Picard 2.20.2 MarkDuplicates (https://broadinstitute.github.io/picard/), and the marked BAM files were used as input for variant calling. The alignment rate, coverage, and mean depth were calculated using Sambamba (Tarasov et al., 2015)

3.3.4. CNV, CNVR and candidate of differentiated gene definition

A combination of the CNVnator v0.4.1 (Abyzov et al., 2011) and LUMPY v0.3.1 (Layer et al., 2014) software was used to identify putative CNV of porcine genomes. CNVnator is a read depth method while LUMPY uses discordant alignment such as split reads and paired-end mapping. CNVs of all samples were called with a bin size of 200 bp by CNVnator and filtered with size (> 1 kb), p-value calculated using t-test statistics (< 0.001) and fraction of reads with zero mapping quality (MQ0 < 0.5). The CNVs in unplaced scaffolds were removed. Structural variations including CNV were detected by 'lumpyexpress' command of LUMPY with default parameter (Layer et al., 2014). Overlapped copy number variable regions with same type of CNV between results of CNVnator and LUMPY were defined as concordant CNVs in every individual. The chromosomal distribution of the concordant CNVs were compared between male and female, p-arm and q-arm, and among populations. A 50% reciprocal overlap between filtered CNVs was defined as copy number variation region (CNVR) using CNVRuler (Kim et al., 2012). CNVRs found in two and more of individuals were used for downstream analysis to minimize false-positive. Copy number of every gene on CNVR were calculated based on aligned read depth and normalized using CNVnator (Abyzov et al., 2011). The normalized copy number of neutral region from diploid autosome was assumed to be 2.0.

3.3.5. Hierarchical clustering based on CNVR

To cluster individuals according to their CNV similarities, I made a vector representing presence or absence of CNV for each individual of genes on CNVRs. Hierarchical clustering with 1000 bootstrap resampling was performed on these vectors for genes on autosomal CNVR using *pvclust* with the default option in R (Suzuki & Shimodaira, 2006). The 'correlation' and 'average' were used as distance measures and the agglomerative method, respectively. The approximately unbiased (AU) *p*-value was calculated by multiscale bootstrap resampling. The bootstrap probability (BP) *p*-value was calculated by ordinary bootstrap resampling based on the unweighted pair-group average method (UPGMA).

3.3.6. Copy number variable genes between populations

The normalized copy number of genes on CNVRs of all individuals was calculated using CNVnator (Abyzov et al., 2011). The normalized copy number of the neutral region from diploid autosome was assumed to be 2.0. V_{ST} of normalized copy number between a pair of populations was calculated as $V_{ST} = (V_T - V_S)/ V_T$, where V_T is the total variance of normalized copy number among all individuals from both populations, and V_S is the average of variance within each population, weighted by the number of individuals in the population (Redon et al., 2006). After excluding the ten populations (ANG, BRI, BUN, GLO, LIN, NES, WJP, WRU, WSW, ZAN) with a single animal V_{ST} between pairs of 56 *Sus scrofa* populations were calculated. Mean V_{ST} of all genes on autosomal CNVRs in each pair of breeds were visualized using *pheatmap* in R (Kolde, 2012). In addition, the V_{ST} of autosomal copy number variable genes were calculated between AD, AW, ED, EW, and NEW. These results were visualized as Manhattan plots using *qqman* package in R (Turner, 2014).

One-way ANOVA test on copy number of every genes on autosomal CNVRs were performed on 5 groups including AD, AW, ED, EW, and NEW. As a *post hoc* test of ANOVA, *Scheffe* test was performed on genes of which ANOVA resulting *p*-values was smaller than 0.05. Genes on CNVR which satisfy both upper 1% pairwise V_{ST} and the *p*-value less than 0.05 of *Scheffe* test after one-way ANOVA were defined as population differentiated genes. Hypothetical, putative, predicted, or uncharacterized genes, as well as pseudo-genes, were excluded.

3.4. Results

3.4.1. Sequence alignment, CNV calling and CNVR definition

The coverage and sequencing depth are important for the credibility of CNVs called using the read depth information of short read alignment. Sequence alignment statistics including mapping rate, coverage and mean depth of all samples were summarized in Table 3.1. In my dataset, the minimum mean depth was higher than 5.06x, and the mean values of alignment rate, coverage, and mean depth of coverage were about 99.3%, 96.4%, and 18.5x, respectively (Table 3.1). Number of CNVs defined by CNVnator, Lumpy and consensus CNV of the two software was summarized in Table 3.2. Lumpy called more CNVs especially deletion than CNVnator in most of individuals. After calling and filtering CNVs, genome-wide CNVRs were identified. Chromosome-wise distribution of CNVs and their total length was summarized in Table 3.3. Among chromosomes, the ratio of total length of CNV to chromosome size were the largest in chromosome Y, followed by chromosome 12 and 6 while the smallest in chromosome 18 followed by 16 and 15. Total length of CNVs were larger in female than male in chromosome 12 and 2, while smaller in chromosome 11, and 16 (Table 3.4). CNV distribution on p-arm and q-arm were also compared based on centromeric region defined in the reference genome. Most of centromere-defined chromosome had more CNVs on q arm while less CNVs on q arm in chromosome 3, 5 and 9 (Table 3.5). Distribution of CNVR larger than 100kb and 500kb were visualized separately in Figure 3.1. Average size of autosomal CNV of AD, AW, ED, EW and NEW were about 51.7, 51.9, 37.9, 26.6 and 9.0 Mbp. Average lengthening and shortening of chromosomal length in each group was summarized in Table 3.6.

Table 3.2. Results of autosomal CNV calling using CNVnator and Lumpy

	Number of duplication			**Number of deletion**		
BioSample	**CNVnator**	**Lumpy**	**Concordant**	**CNVnator**	**Lumpy**	**Concordant**
SAMEA3497824	309	708	451	1975	19548	1253
SAMN04440479	468	22896	3471	1342	7043	5188
SAMN06348392	1479	1183	1633	1312	14805	1554
SAMN06348393	738	1133	1236	1014	16559	1235
SAMN06348414	528	1283	1580	1193	18952	1814
SAMN06348415	544	1138	1074	1054	16000	1153
SAMN06348416	614	1204	1011	1237	15134	1418
SAMN06348417	732	1327	1428	1086	17578	1363
SAMN06349454	571	1215	1080	1138	17978	1741
SAMN06349455	542	1157	995	1232	18622	1386
SAMN06349456	544	1207	1035	1083	18319	1331
SAMN06349457	502	1187	868	1407	18373	1532
SAMN06349458	812	1249	1449	1335	15855	1665
SAMN06349459	527	1237	981	1352	17163	1581
SAMEA3497827	299	673	367	2064	20901	1011
SAMN03566761	421	299	159	621	2865	296
SAMN04440475	409	18342	2983	963	5291	3350

SAMEA3497828	349	637	495	1764	19953	1043
SAMN03566754	915	562	382	837	4679	515
SAMN03566755	647	521	392	807	5697	683
SAMN03566756	502	640	493	855	6505	654
SAMN03566757	587	738	568	1883	8956	1354
SAMN03566758	410	581	514	732	6135	524
SAMN03566759	1071	599	896	2646	6224	1525
SAMN03566760	536	636	485	888	6091	655
SAMN02298127	511	1901	2086	4530	37357	8043
SAMN02298128	477	1916	2134	3850	36919	8080
SAMN02298129	494	1858	2124	2822	30807	5073
SAMN02298130	536	1858	2076	3679	37418	6181
SAMN02298131	456	1890	2126	3734	35593	5369
SAMN02298132	449	1888	2175	3985	37873	7002
SAMEA3497830	333	685	488	1615	20234	840
SAMEA3497826	348	842	564	1944	22782	1595
SAMEA3497832	314	690	531	832	18374	517
SAMEA3497835	329	486	265	1124	4504	503
SAMEA3497836	439	425	404	1014	6660	506
SAMEA3497837	302	608	399	1584	16536	987
SAMEA3497838	317	480	289	5074	16320	2338

SAMEA3497839	608	252	235	2329	7321	389
SAMEA3497840	2825	240	952	3812	7691	848
SAMN00005058	255	226	52	605	6669	143
SAMN03031126	540	298	244	806	5434	209
SAMN03031127	506	644	450	1003	9164	498
SAMN03031128	1008	542	481	1233	9064	708
SAMN09930402	789	901	858	1074	13120	758
SAMN09930403	709	1163	1321	1679	14511	1950
SAMN12122743	770	1205	1203	1201	14503	1209
SAMN12122744	688	1162	1356	1676	14513	1923
SAMN28745316	360	1453	1546	990	3717	3107
SAMN28745317	368	1425	1536	1077	3944	3156
SAMN28745318	332	1245	1125	951	3440	2039
SAMN28745319	344	1032	990	862	3001	1793
SAMN28745320	395	1542	1530	994	4006	2948
SAMN02298079	567	1749	2244	4387	36881	7781
SAMN02298080	469	1875	1910	3679	36828	5973
SAMN09930385	535	1702	1369	3113	23586	4339
SAMN09930386	535	1762	1421	3113	23497	5570
SAMN09930387	535	1701	1468	3113	23726	4749
SAMN09930388	671	1678	1791	2003	23311	3458

SAMN09930389	544	1729	1911	3188	23311	4844
SAMN09930390	643	1720	2243	2919	23990	3891
SAMN09930391	648	1588	1933	3125	25063	5111
SAMN09930392	541	1701	1681	3280	22212	5120
SAMN09930393	535	1625	1864	3113	26188	5674
SAMN09930394	535	1720	1868	3113	28439	5423
SAMN09930395	616	1783	2185	2899	27019	5347
SAMN09930396	670	1533	1907	3328	25691	5826
SAMN09930397	540	1689	2046	3064	25623	4812
SAMN09930398	543	1595	1881	3132	24631	4790
SAMN09930399	609	1675	1807	3129	21430	5464
SAMN09930400	640	1662	1738	3087	23874	5039
SAMN09930401	494	1695	1642	3275	26309	5351
SAMN12122745	586	1740	2436	4401	36871	7859
SAMN12122746	467	1875	1871	3654	36825	5953
SAMN04538376	8590	747	3474	1779	7075	638
SAMN04538598	8604	716	2720	2043	6919	1155
SAMN04538599	8297	542	4313	2259	5382	802
SAMEA3497842	9563	657	4309	5090	15965	2315
SAMEA3497843	1299	397	606	808	5976	384
SAMEA3497844	6557	463	1292	2574	3911	800

SAMN04440474	409	20375	2385	967	5173	2883
SAMN02298115	482	1588	1609	6058	35653	6053
SAMN02298116	509	1578	1491	5189	35345	5783
SAMN02298117	422	1356	1023	6028	34095	6124
SAMN02298118	429	1635	1348	5116	35772	5422
SAMN02298119	478	1633	1460	5497	35983	6532
SAMN02298120	485	1388	1113	8346	35981	9672
SAMN02904857	348	516	407	6279	11955	2079
SAMN03421607	327	603	322	778	4915	611
SAMN05362554	334	561	303	778	3416	527
SAMN06895012	333	538	331	770	3986	457
SAMN06349462	470	1180	1069	1286	16935	1773
SAMN06349463	507	1378	1281	1657	18773	2152
SAMN06349464	458	1176	983	1389	17010	1563
SAMN06349465	569	1313	910	1228	17774	1595
SAMN06349466	511	1094	952	1747	16347	1956
SAMN06349467	812	1312	1305	1457	18549	1862
SAMEA3497793	565	627	710	974	10997	1141
SAMEA3497794	431	612	385	1400	9028	1244
SAMN04440480	380	468	246	1335	5169	405
SAMEA3497795	1095	658	825	5291	18035	3884

SAMEA3497796	424	466	280	1440	8741	528
SAMEA3497797	399	504	270	1155	12142	613
SAMN28745321	413	1507	1678	1150	4397	2625
SAMN28745322	404	2101	2176	1296	5165	5414
SAMN28745323	375	1465	1310	2591	4231	4780
SAMN28745324	390	2006	2027	1257	5145	5183
SAMN28745325	374	1591	1541	1230	4450	3615
SAMN28745312	362	1686	1663	1187	4436	3320
SAMN28745313	399	1194	1064	914	3719	2037
SAMN28745314	379	1609	1660	1098	4106	3166
SAMN28745315	351	1277	1194	935	3641	2249
SAMN28745297	345	1158	885	1317	3656	2582
SAMEA3497847	1299	443	843	2603	11263	1341
SAMEA3497850	408	399	288	3460	10079	921
SAMEA3497851	329	513	293	989	7747	410
SAMN04440476	400	15923	2685	998	5084	2590
SAMEA3497798	385	719	489	1478	13674	1084
SAMEA3497799	641	640	708	1856	8283	1113
SAMEA3497852	301	655	388	5405	22087	1749
SAMN02298087	915	1393	1943	5614	34190	10353
SAMN02298088	587	1753	1835	5114	37796	11556

SAMN02298089	639	1824	2358	3226	41405	8280
SAMN02298090	580	1865	2262	2765	32966	5476
SAMN02298091	996	1815	2779	3005	40922	9396
SAMN02298092	1728	1692	4218	3811	35355	9827
SAMN02298133	374	1707	1272	3368	37418	4422
SAMN02298134	416	1652	1276	2927	33553	4887
SAMN02298135	453	1628	1755	3037	33056	5909
SAMN02298136	486	1645	1524	3091	32445	4205
SAMN02298137	1745	1698	3987	2475	30895	5780
SAMN02298138	645	1669	1925	2692	35514	4968
SAMEA3497854	311	442	210	1098	7690	508
SAMEA3497855	337	490	262	1004	8712	506
SAMN02665304	503	239	196	2205	1535	274
SAMN02665305	347	569	388	1710	10208	999
SAMN02665306	389	583	356	2362	9952	761
SAMEA3497800	1078	653	1159	1014	13774	1325
SAMN04440481	445	689	492	1278	10794	740
SAMEA3497801	428	759	652	1059	11785	1021
SAMEA3497802	511	703	559	3107	12945	1392
SAMEA3497803	576	703	635	4565	17333	2263
SAMEA3497804	408	518	323	1193	11909	646

SAMEA3497805	444	693	598	966	13939	849
SAMEA3497806	409	755	615	1021	9810	960
SAMEA3497807	8015	588	776	651	7862	386
SAMEA3497808	399	685	447	1007	12663	957
SAMEA3497809	745	923	1008	6055	20231	6744
SAMEA3497856	293	707	528	2226	23276	1053
SAMN02298121	556	1575	2053	4806	34117	6974
SAMN02298122	442	1595	1380	5011	33553	6286
SAMN02298123	657	1532	1587	4372	26085	7581
SAMN02298124	472	1488	1236	4645	30250	6701
SAMN02298125	524	1735	2088	4254	32112	6737
SAMN02298126	398	1643	1454	1882	35387	3867
SAMN01894448	5518	220	944	5168	5205	469
SAMN06393132	627	1361	1790	1309	19924	1921
SAMN06393133	633	1394	1400	1151	19731	1570
SAMN06393134	522	1334	1178	1387	19837	1887
SAMN06393485	622	1252	1044	1238	17358	1505
SAMN06394064	665	1284	1195	1288	17109	1636
SAMN06394627	647	1277	1382	1718	19044	2388
SAMN08035066	400	1368	1202	1099	4935	2219
SAMEA3376934	363	256	156	1238	6123	277

SAMEA3376936	825	511	456	2774	12435	781
SAMEA3376937	393	646	393	5766	21339	3133
SAMEA3376938	388	683	656	7321	21317	3684
SAMEA3376939	338	799	557	2464	22983	1267
SAMEA3376940	19101	582	6683	874	15441	573
SAMEA3376941	6116	411	1212	810	2260	61
SAMEA3376942	414	607	542	1162	18861	776
SAMEA3376943	409	596	317	1076	17543	674
SAMEA3376944	32725	634	18549	20518	18584	6575
SAMEA3497791	4215	256	1938	2460	2437	488
SAMEA3497860	4114	129	79	280	490	3
SAMN04440477	411	19591	2153	955	5078	2315
SAMN02460623	616	329	221	940	8597	248
SAMN02460625	719	330	321	1043	8688	248
SAMN02460626	893	325	253	913	3178	361
SAMN02460627	492	205	166	1118	3781	408
SAMN03331745	332	241	117	851	6445	193
SAMN04440482	472	473	397	1329	6985	476
SAMEA3497862	327	485	299	1106	8667	450
SAMEA3497863	337	559	381	1084	8225	580
SAMN02646543	3470	310	449	5610	12662	1861

SAMN02646544	2862	240	889	4347	10387	1329
SAMN02646545	2151	345	741	2003	14124	972
SAMEA3497810	371	641	482	1005	15234	876
SAMEA3497811	373	681	400	1084	12848	888
SAMEA3497864	385	389	206	2271	7618	607
SAMEA3497865	424	479	250	2396	9736	1227
SAMEA3497866	825	410	239	2613	10088	1141
SAMEA3497867	370	646	365	3949	18529	2100
SAMEA3497868	1130	487	392	1672	3992	557
SAMEA3497869	640	410	220	973	2143	481
SAMEA3497870	1945	89	44	478	323	26
SAMEA3497871	625	381	167	3408	10126	1519
SAMEA3497872	380	372	286	2710	8500	821
SAMEA3497873	373	349	141	2081	7900	466
SAMEA3497874	326	522	208	2215	7386	1142
SAMEA3497875	482	498	409	1275	9568	576
SAMEA3497876	1628	493	669	1708	9605	693
SAMEA3497879	612	605	517	3386	16261	1600
SAMEA3497886	338	628	461	1070	7963	710
SAMEA3497887	357	606	400	1347	8505	975
SAMEA3497888	1281	543	935	1764	5689	1327

SAMEA3497823	1888	761	1858	5180	17944	4374
SAMN03031171	1387	652	1459	5694	20080	3566
SAMN03031172	1956	287	448	1203	10191	351
SAMN03031173	1518	196	347	6984	11134	890
SAMN03031174	524	253	118	5700	10929	1265
SAMN03031175	1198	290	419	6671	12974	1824
SAMN03031176	800	249	129	4108	9891	1146
SAMN03031177	2888	262	209	1236	9743	239
SAMN03031178	1037	773	1249	2755	19752	1900
SAMN03031179	920	233	132	1262	9346	354
SAMN03031180	1100	613	871	4978	19462	1799
SAMEA3497821	1063	707	1091	4089	19764	2766
SAMEA3497822	1260	377	439	953	2966	403
SAMEA3497884	456	549	280	1704	10349	824
SAMEA3497885	345	572	333	963	9645	564
SAMN05362551	5510	294	3599	5199	4575	749
SAMEA3497815	4126	139	93	333	735	4
SAMEA3497816	861	709	911	3285	19194	1457
SAMEA3497818	465	1933	2039	1474	18433	3466
SAMEA3497819	413	903	703	1124	18023	1355
SAMN01894459	412	203	188	4298	9754	1065

SAMN02298081	455	1927	1871	3268	37978	5708
SAMN02298082	472	2001	1846	3183	43698	5363
SAMN02298083	510	1845	2047	3074	45765	7894
SAMN02298084	916	1398	2284	3284	40259	6833
SAMN02298085	1625	1321	2934	5139	43197	10825
SAMN02298086	1700	1378	3077	4941	41497	10432
SAMN02904855	372	556	417	755	3114	462
SAMN05362552	336	465	204	693	2816	330
SAMEA3497877	668	351	190	1571	5090	479
SAMN02298093	451	1662	1517	3439	37833	6425
SAMN02298094	450	1718	1814	4560	40323	7371
SAMN02298095	1089	1387	1896	3508	38247	7028
SAMN02298096	413	1280	850	3302	39633	4751
SAMN02298097	467	1374	1110	3678	41100	5701
SAMN02298098	466	1883	1712	2371	38507	5559
SAMN12122795	525	1089	903	1048	17274	1178
SAMN12122796	623	1443	1388	1137	19027	1595
SAMN12122797	785	1270	1420	1215	18409	1430
SAMN12122798	604	1218	1165	1056	17611	1460
SAMN12122799	634	1378	1361	1603	19526	1914
SAMN12122800	630	1168	966	2572	18717	2301

SAMN01894407	2424	238	691	4070	6219	1133
SAMN02298111	556	1621	1724	6518	36116	11373
SAMN02298112	521	1493	1213	6708	35670	8833
SAMN02298113	580	1529	1586	5609	34227	9489
SAMN02298114	583	1534	1463	7024	34083	7763
SAMN12122783	628	1496	1661	1261	20230	2136
SAMN12122784	550	1414	1268	1507	19715	1946
SAMN12122785	603	1181	1463	1645	16628	1786
SAMN12122786	508	1121	1219	1366	15590	1338
SAMN12122787	481	1248	1189	1141	17024	1373
SAMN12122788	498	1225	1036	1865	18234	2234
SAMN01894388	619	339	237	3933	3407	796
SAMN01894391	1574	244	522	4632	3494	1205
SAMN01894434	873	301	370	5115	7792	773
SAMN01894436	1977	231	522	5709	7270	846
SAMN02298105	504	1766	1691	5018	39860	7411
SAMN02298106	656	1711	1746	4644	37198	7380
SAMN02298107	538	1842	1998	4031	35267	6793
SAMN02298108	550	1508	1557	3061	33862	3804
SAMN02298109	604	1988	2199	3131	37058	5641
SAMN02298110	524	1861	2124	5111	35831	8440

SAMN12122789	492	1096	893	1645	18075	1588
SAMN12122790	546	1311	1227	1335	19288	1847
SAMN12122791	625	1319	1458	1241	19145	1561
SAMN12122792	641	1337	1515	1208	19684	1798
SAMN12122793	525	1089	903	1048	17274	1178
SAMN12122794	643	1272	1750	1803	18606	2408
SAMN01894367	1184	299	700	2671	5073	673
SAMN01894370	1741	271	892	2693	4401	778
SAMN02298099	600	1856	1950	3514	36353	6239
SAMN02298100	996	1420	1917	3433	23214	3827
SAMN02298101	470	1910	1567	2533	40498	4296
SAMN02298103	504	1804	1978	2140	39958	5058
SAMN02298104	756	1815	2007	2062	39177	4642
SAMN12122801	537	1251	1185	1104	18227	1560
SAMN12122802	555	1386	1232	1296	20681	1619
SAMN12122803	567	1186	1232	1581	19162	1545
SAMN12122804	558	1227	1213	1429	18370	1671
SAMN12122805	751	1162	1354	2168	18923	2061
SAMN12122806	705	1257	1200	1186	19012	1695
SAMEA3497812	563	668	684	2843	26370	1982
SAMEA3497813	537	638	555	3034	25654	1803

SAMEA3497853	13315	487	6462	26654	12153	6190
SAMN04440478	422	20780	2509	1020	5325	2967
SAMN12122747	610	1136	1094	2339	18017	2149
SAMN12122749	712	1106	1532	2373	16002	2484
SAMN12122750	897	1103	1595	2383	13550	2570
SAMN12122751	791	1085	1406	2255	14154	2068
SAMN12122752	750	1195	1345	2293	16328	2384
SAMN12122753	629	1135	1156	2453	16652	2401
SAMN12122754	1198	1034	1828	2860	14991	2475
SAMN12122755	963	1050	1553	2353	16888	2846
SAMN12122756	714	1144	1557	2336	16127	2586
SAMN12122757	897	1071	1535	2523	14221	2573
SAMN12122758	553	1203	1065	2234	15849	1879
SAMN12122759	804	1112	1395	2101	4934	1891
SAMN12122760	771	1022	1139	2601	17072	2728
SAMN12122761	931	1121	1346	2320	13321	2058
SAMN12122762	1189	1062	1560	2073	14441	2276
SAMN12122763	1282	825	1432	2531	12925	1842
SAMN12122764	558	1096	1348	2383	15129	2749
SAMN12122765	836	1095	1319	2364	15184	2491
SAMN12122766	639	1165	1091	2201	14912	2312

SAMN12122767	697	1093	1223	2425	13468	2137
SAMN12122768	803	1005	1238	2521	15000	2431
SAMN12122769	707	1138	1395	2029	9663	1694
SAMN12122770	908	1052	1248	2296	13896	2534
SAMN12122771	786	994	1099	2065	3787	1425
SAMN12122772	827	1414	1961	2799	14820	3754
SAMN12122773	687	1216	1497	2238	14902	2645
SAMN12122774	1098	1007	1406	2401	11833	2133
SAMN12122775	825	1226	1736	2268	15309	2293
SAMN12122776	6435	1073	6923	5225	12753	4017
SAMN12122777	1512	1013	1644	2543	12664	2434
SAMN12122778	1434	891	1362	3248	14154	3284
SAMN12122779	600	1221	1221	2860	15882	3548
SAMN12122780	523	1110	1006	2402	15181	2212
SAMN12122781	1011	1058	1336	2339	14836	2470
SAMN12122782	566	1182	1157	2457	17576	2773
SAMEA3497814	382	697	542	848	13806	747
Total	362083	447792	420422	850187	5647525	899863
Average	1103.9	1365.2	1281.8	2592.0	17218.1	2743.5

Table 3.3. Chromosome-wise distribution of CNVs

Chr.	Chromosome size	Average Count	Average total length of CNVs (bp)	CNV count /Chromosome size	Total CNV length /Chromosome size
1	274330532	276.9	5.6.E+06	1.0.E-06	2.0.E-02
2	151935994	192.8	3.1.E+06	1.3.E-06	2.1.E-02
3	132848913	129.3	1.9.E+06	9.7.E-07	1.4.E-02
4	130910915	103.8	1.5.E+06	7.9.E-07	1.1.E-02
5	104526007	135.2	1.6.E+06	1.3.E-06	1.5.E-02
6	170843587	278.7	6.8.E+06	1.6.E-06	4.0.E-02
7	121844099	134.9	2.0.E+06	1.1.E-06	1.6.E-02
8	138966237	121.3	2.0.E+06	8.7.E-07	1.4.E-02
9	139512083	188.6	3.5.E+06	1.4.E-06	2.5.E-02
10	69359453	69.5	1.2.E+06	1.0.E-06	1.8.E-02
11	79169978	104.2	2.0.E+06	1.3.E-06	2.6.E-02
12	61602749	187.6	4.3.E+06	3.0.E-06	7.0.E-02
13	208334590	172.7	3.5.E+06	8.3.E-07	1.7.E-02
14	141755446	162.2	2.8.E+06	1.1.E-06	2.0.E-02
15	140412725	89.9	1.5.E+06	6.4.E-07	1.0.E-02
16	79944280	47.8	7.3.E+05	6.0.E-07	9.2.E-03
17	63494081	62.5	9.7.E+05	9.8.E-07	1.5.E-02
18	55982971	35.1	3.2.E+05	6.3.E-07	5.7.E-03
X	125939595	168.9	2.1.E+06	1.3.E-06	1.7.E-02
Y	43547828	130.5	2.3.E+07	3.0.E-06	5.3.E-01

Table 3.4. Different distribution of chromosome-wise CNV between sexes

Chromosome	Average Count		Average Length	
	Female	**Male**	**Female**	**Male**
1	279.0	275.5	4957364.1	5971559.6
2	223.9	172.4	3513204.7	2879474.2
3	152.9	113.8	1999578.3	1877495.1
4	109.4	100.1	1329709.6	1530352.7
5	151.8	124.2	1642780.2	1558191.9
6	320.4	251.3	7602342.3	6352111.3
7	152.5	123.4	2114238.4	1852956.5
8	132.2	114.2	1995790.9	1944930.2
9	193.8	185.2	3346537.5	3674041.2
10	79.5	63.0	1343991.6	1157732.3
11	97.5	108.6	1482468.5	2422472.4
12	222.5	164.6	4891888.6	3923611.0
13	176.0	170.5	3263217.5	3682771.7
14	181.0	149.9	2786934.5	2814254.7
15	93.7	87.3	1442594.0	1464966.6
16	47.1	48.2	549886.0	855990.8
17	68.3	58.7	951341.7	988677.6
18	40.5	31.6	317290.6	321734.8

Table 3.5. CNV distribution on p-arm and q-arm

Chr	Size	Position of centromere	Centromeric region Start	Centromeric region End	Count of CNVs on p arm	Length of CNVs on p arm	Count of CNVs on q arm	Total length of CNVs on q arm
1	274330532	Metacentric	92615481	93430514	22979	407979179	65615	1363551600
2	151935994	Metacentric	50550173	50777308	24053	298310579	38931	726138451
3	132848913	Metacentric	41776737	41860603	25925	445729090	16363	183645641
4	130910915	Metacentric	46443460	46472085	11512	133812320	22349	341183976
5	104526007	Metacentric	39774025	40207105	19450	205052779	9569	89498852
6	170843587	Metacentric	38712705	38886534	14141	207824957	77250	2038089982
7	121844099	Metacentric	24578125	24606761	8376	89727180	35703	551440537
8	138966237	Metacentric	54585508	54685241	16460	229281427	22775	393627036
9	139512083	Metacentric	63144551	63503859	31777	604691204	17730	436791618
11	79169978	Metacentric	11220831	11222126	5544	58812769	28625	613557683
11	79169978	Metacentric	35726738	35878206	15575	228520441	18382	433119479
13	20833459	Acrocentri	34	152474	0	0	56594	1153180448

	0	c						
15	140412725	Acrocentric	1649	36105	7	117512	29434	477213408
15	140412725	Acrocentric	56407100	56427869	9444	131183914	20029	346416683
17	63494081	Acrocentric	63189675	63361433	490	2590346	19990	316653604
18	55982971	Acrocentric	619	17212	0	0	11518	104951256
Y	43547828	Metacentric	42496777	42515903	4813	116874562	21329	4501735448

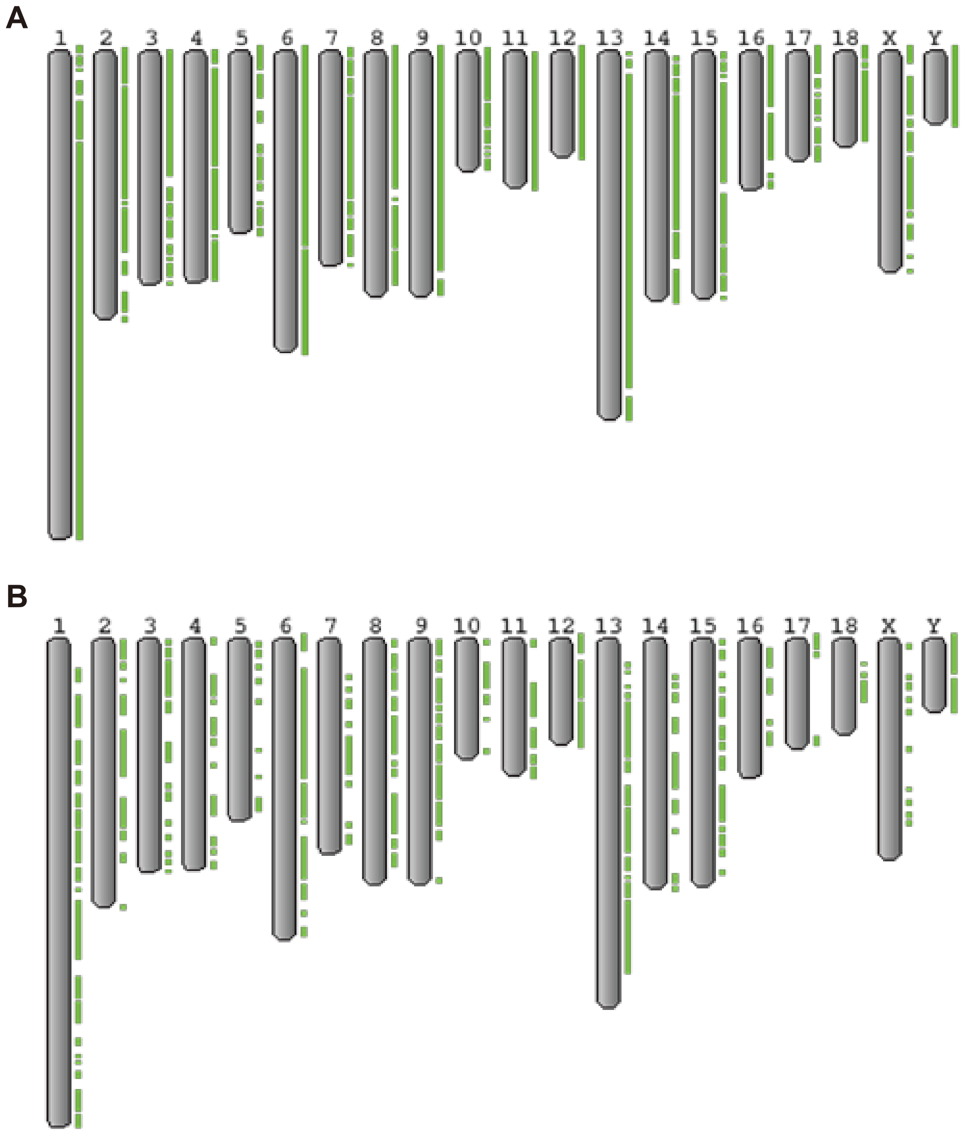

Figure 3.1. CNVR distribution

Distribution of CNVRs larger than 100 kb (A) and 500 kb (B) were visualized separately. Green rectangles on the right side of chromosomes represents CNVRs.

Table 3.6. Average lengthening and shortening of chromosomal length in each groups

Chr.	ED	AD	EW	AW	NEW
1	3214670	1168083	2919576	1921420	16679
2	-946546	-1495163	-1451222	-1098171	-570499
3	-65148	-925775	-1060788	-1033091	-81260
4	-246122	-143454	195848	-547591	-120736
5	-9395	518972	437522	34392	-297349
6	-3848416	-1263151	-4566279	195915	-21751
7	541222	343037	720710	676843	703493
8	1285469	391771	37103	-145925	1804538
9	381317	1676882	411754	996183	150019
10	-714424	-416789	-551074	-619322	-503007
11	-447230	-399562	-606681	-929036	-236927
12	-3894106	-3811417	-3275702	-2108478	-20098
13	1005408	1194139	560586	901623	-70915
14	-221560	-508014	795413	-441246	-17301
15	53249	-220396	62944	130182	-272414
16	-85169	263877	435901	245735	-107509
17	-240684	-16081	-176652	-487280	-555816
18	-5873	-141618	-40658	7952	55551
X	-20427823	-11640637	-3874057	-3690100	-540149
Y	-26411562	-22957742	-28480105	-23008739	-29302628

There were population specific lengthening and shortening of chromosomal length in chromosome 4-6, 8 and 14-18 (Table 3.6).

3.4.2. Population differentiation based on copy number variable genes

Hierarchical clustering of all individuals was performed on vectors considering the presence or absence of autosomal CNVRs (Figure 3.2). Mean pairwise V_{ST} values of breeds including more than one animal were calculated on 315 animals from 43 populations and visualized as a heatmap with hierarchical clustering (Figure 3.3). The V_{ST} range is from 0 to 1, with a higher value indicating a larger difference. The pairwise mean V_{ST} values of five groups were as following: AD-ED, 0.009; AD-AW, 0.032; AD-EW, 0.015; AD-NEW, 0.005; ED-EW, 0.012; ED-AW, 0.040; ED-NEW, 0.005; AW-EW, 0.020; AW-NEW, 0.007; EW-NEW, 0.020. The average of the pairwise V_{ST} in groups was about 0.017, and the average in breed level was 0.240.

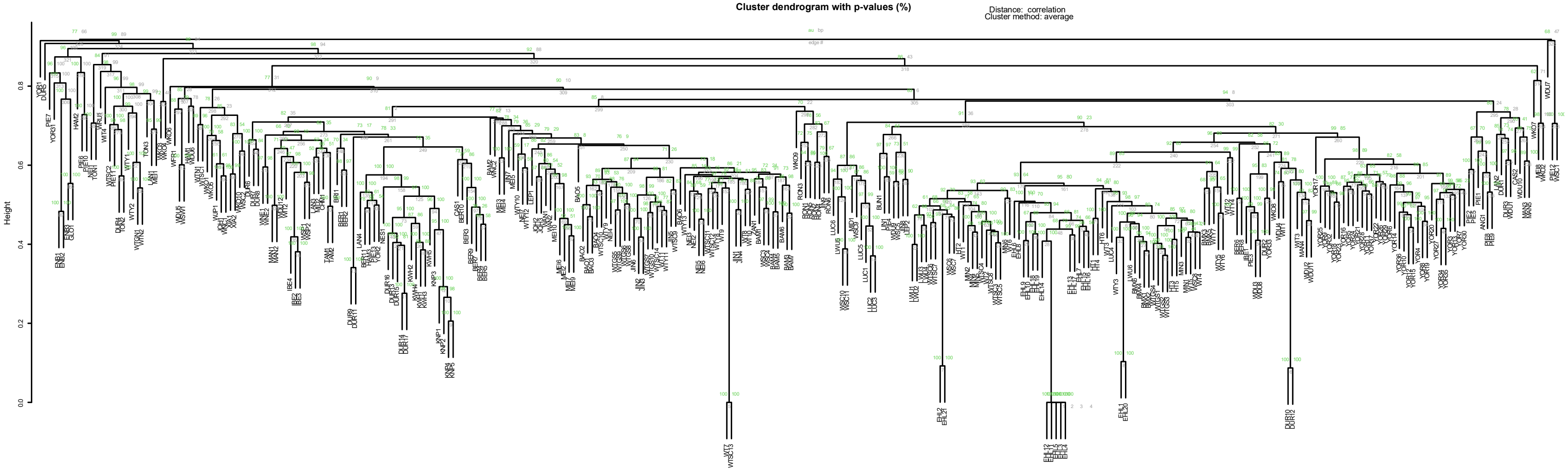

Figure 3.2. Hierarchical clustering tree

For every individual, the absence or presence of CNVs in autosomal CNVRs was converted to a vector made of ‘0’s and ‘1’s. The hierarchical clustering was performed on these vectors representing each individual. The bootstrap value was shown under the edges of the clustering. The ㅈ approximately unbiased (AU) and the bootstrap probability (BP) p-value were written in red and green letters on the edges after multiplied by 100.

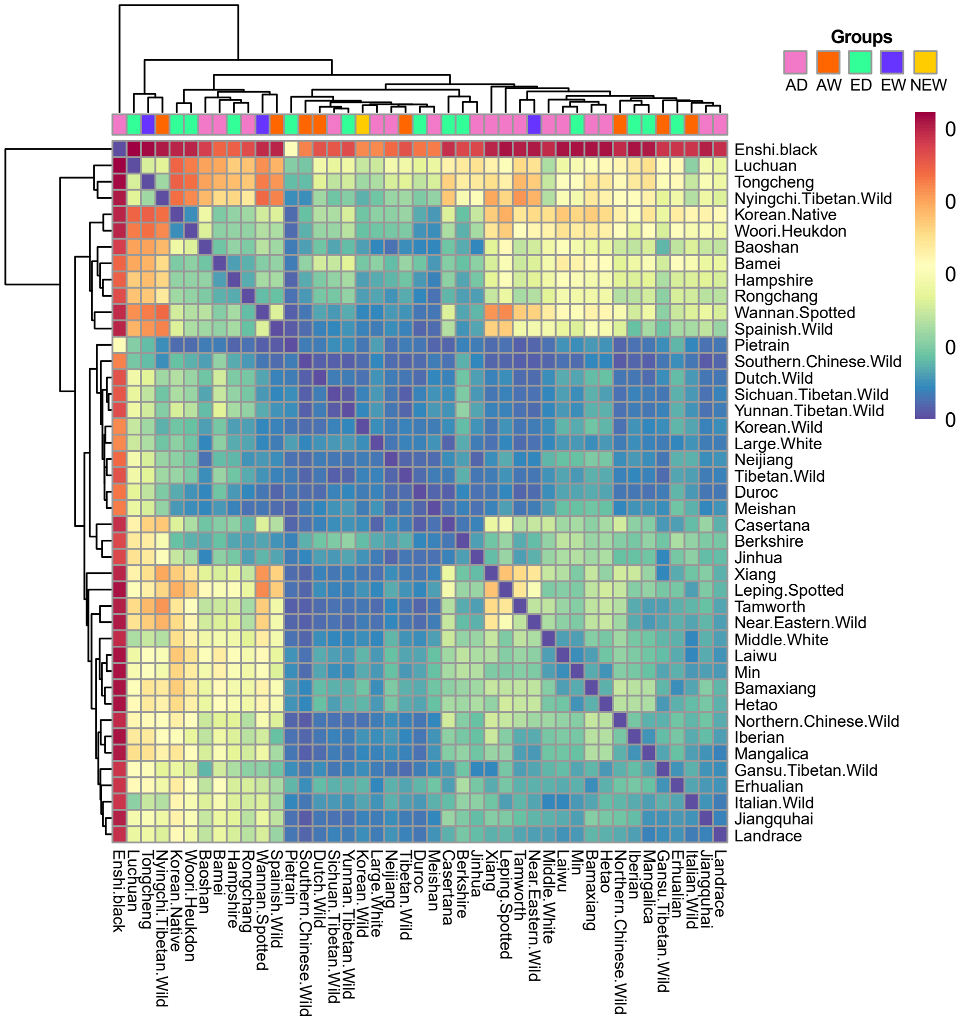

Figure 3.3. Heatmap representing average of pairwise V_{ST} between breeds

Average of pairwise V_{ST} of genes on autosomal CNVRs were calculated between all pairs of breeds which included more than 1 sample. Clustering was performed only on the mean pairwise V_{ST}.

3.4.3. Copy number variable genes across populations

Candidates of copy number variable genes were suggested based on the two criteria; pairwise V_{ST} and one-way ANOVA across five groups, including AD, ED, AW, EW, and NEW. First, V_{ST} was calculated between pairs of five groups. The upper 1% and upper 0.1% values of pairwise V_{ST} between groups were about 0.159 and 0.409, respectively. Pairwise V_{ST} of genes on autosomal CNVR were visualized as Manhattan plot (Figure 3.4). There were some peaks shared by pairs of groups. I suggested the shared peaks between pairs including a same group as the regions with copy numbers distinct from other groups.

Then, differences of normalized copy numbers across the five groups were tested using the one-way ANOVA followed by *Scheffe* test. Among genes of which the *p*-value was below 0.05, 111 genes of which V_{ST} values in the upper 0.1% of at least a pair of groups defined as copy number variable genes. 15 genes were remained after excluding hypothetical, putative, predicted, or uncharacterized genes, as well as pseudo-genes (Table 3.7). Among these copy number differentiated genes, group-wise average copy number of every 1kb of *EEA1* were visualized in Figure 3.5.

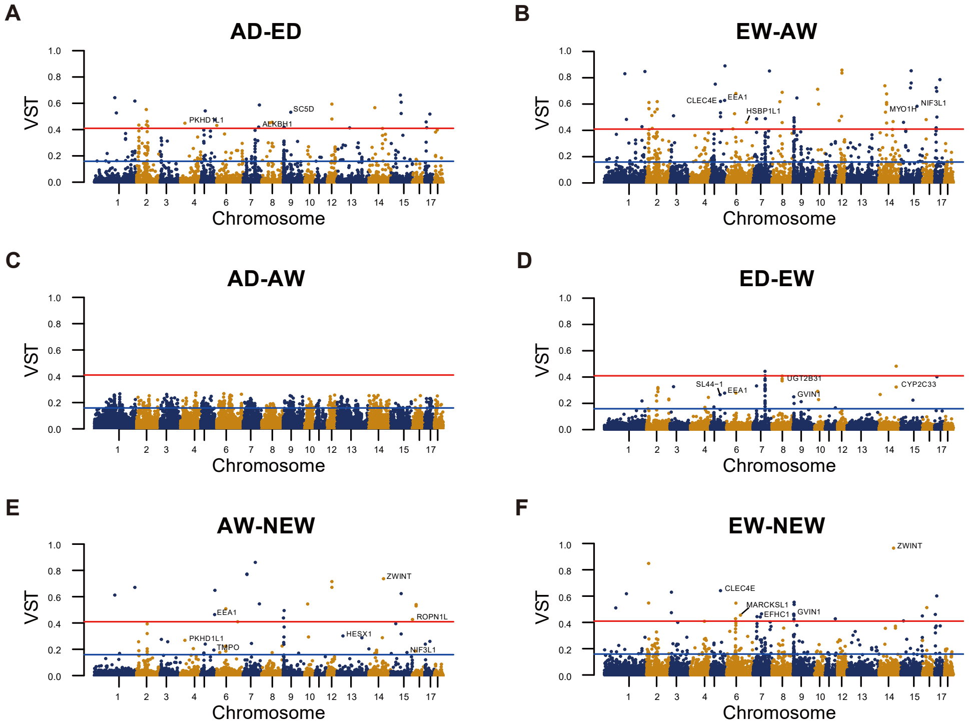

Figure 3.4. Manhattan plot of V_{ST}

V_{ST} of genes on autosomal CNVRs were visualized as Manhattan plots. The center point of genes was used as an x-coordinate value. Genes with significantly different pairwise V_{ST} in upper 0.1% were marked by their names. Name of hypothetical, putative, predicted or uncharacterized genes and pseudo-genes were excluded due to lack of space. The upper 1% percentile V_{ST}, 0.157, and upper 0.1% percentile, 0. 409, were shown as blue and red lines, respectively.

Table 3.7. Genes with differentiated copy number between populations

Gene	Chr.	Start	End	ANOVA p-value	Scheffe p-value	VST upper 0.1% pair	Average copy number				
							AD	ED	AW	EW	NEW
PKHD1L1	4	28023448	28176331	2.74.E-41	AD-ED,AD-EW,AD-NEW,AW-ED,AW-EW,AW-NEW	AD-ED	1.5	1.5	1.8	1.8	2.0
CLEC4E	5	63219229	63228566	2.05.E-42	AD-ED,AD-EW,AW-ED,AW-EW,ED-EW	AW-ED,AD-EW,AW-EW,EW-NEW	0.9	0.9	1.6	1.8	1.0
EEA1	5	90131707	90257014	5.01.E-38	AD-ED,AD-EW,AD-NEW,AW-ED,AW-EW,AW-NEW,ED-EW,ED-NEW	AD-EW,AW-EW,AW-NEW	9.1	8.8	10.9	14.8	17.3
MARCKSL1	6	88785412	88787772	9.60.E-14	AD-ED,AD-EW,AW-ED,AW-EW	EW-NEW	1.9	1.7	1.3	1.1	2.3
HSBP1L1	6	127960330	127972241	1.26.E-26	AD-ED,AD-EW,AW-ED,AW-EW,AW-NEW	AW-EW	2.6	2.7	2.3	2.1	2.0
EFHC1	7	46244915	46320261	1.64.E-04	AD-NEW,AW-ED,ED-NEW,EW-NEW	EW-NEW	2.0	2.1	2.0	2.0	2.4
ALKBH1	7	100676778	100710414	1.32.E-36	AD-ED,AD-EW,AW-ED,AW-EW	AD-ED	1.9	1.9	2.1	2.1	2.0
UGT2B31	8	66310697	66323755	5.87.E-20	AD-AW,AD-ED,AD-EW,AD-NEW,AW-EW,AW-NEW,ED-EW,ED-NEW	AD-EW	2.6	3.1	3.1	4.5	5.0
GVIN1	9	2874233	2882380	3.34.E-05	AD-EW,ED-EW	EW-NEW	1.1	1.3	1.0	1.8	0.4
SC5D	9	48357115	48372391	9.30.E-51	AD-ED,AD-EW,AW-ED,AW-EW	AD-ED,AW-ED	2.0	2.0	2.4	2.3	2.0
MYO1H	14	41469191	41588934	4.85.E-28	AD-ED,AD-EW,AW-ED,AW-EW,ED-EW	AW-EW	1.7	1.8	1.6	1.5	1.5

ZWINT	14	94094781	94109447	1.66.E-40	AD-NEW,AW-NEW,ED-NEW,EW-NEW	ED-NEW,AD-NEW,EW-NEW,AW-NEW	2.1	2.2	2.1	2.2	4.1
CYP2C36	14	106184665	106219631	2.29.E-27	AD-ED,AD-EW,AW-ED,AW-EW,ED-NEW	AW-ED	3.4	3.7	2.2	2.5	4.4
NIF3L1	15	104583736	104606507	3.01.E-18	AD-ED,AD-EW,AW-ED,AW-EW,ED-EW	AD-EW,AW-EW	1.7	1.8	2.0	2.3	2.2
ROPN1L	16	43299	58569	1.76.E-07	AD-NEW,AW-NEW,ED-NEW,EW-NEW	AW-NEW	1.9	1.8	1.9	2.0	3.0

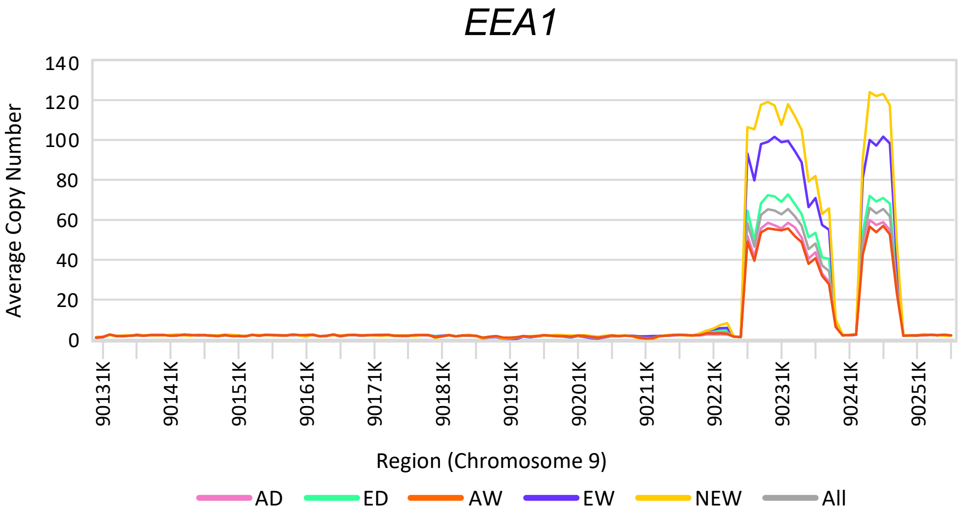

Figure 3.5. Average copy number of 5 groups in *EEA1*

Average copy number around *EEA1* coding region. X-axis indicated genomic region and y-axis indicated average copy number in each group. *EEA1* located from 90,131,707 to 90,257,014 in chromosome 5 and the average copy number of every 1000bp regions from 90,131,001 to 90,258,000 were visualized as a line graph. The two peak regions were 90,227,001–90,240,000 and 90,244,001–90250000.

3.5. Discussion

Since the colonization of wild boar across mainland Eurasia and North Africa within two Mya and domestication started 10,000 years ago, *Sus scrofa* has been adapted to various environments and human needs. In addition to selection pressure, demographic events such as the bottleneck in the last glacial period about 20,000 years ago and migration following farmers intensified the development of various pig breeds. Furthermore, modern breeding programs have accelerated genomic studies on pigs with the aim of improving their value as a source of meat and model animals. In particular, porcine CNV has been a great subject for studying phenotypic variance, especially in quantitative traits, as it can alter gene dose and expression. my study analyzed the largest number of Eurasian wild boar and domesticated pigs with two values to measure the differences in copy number between populations. The first was V_{ST} based on variance, and the second was the one-way ANOVA test. Considering both values together, I present the copy number variable regions and compare the copy number between populations. Chromosome-wise distribution of CNVs were compared by population, sex and chromosomal location such as p-arm and q arm separately. The autosomal CNVs covered larger regions in Asian pigs than European pigs which might be results of reference bias of using single reference representing Duroc. On the other hand, I could not observe any consistent effects of sex and chromosomal location on prevalence of CNVs. There would be other multiple genomic features which affect the probability of CNV occurrence.

Hierarchical clustering was performed on vectors representing the presence or absence of CNVs on autosomal CNVRs. Some individuals were clustered following their groups while others were not. For example, Pietrain individuals were clustered discordant with their breeds. Actually, variance of copy numbers was highest in Pietrain among breeds with the value (1.06) significantly higher than others, followed by the variance of Meishan (0.33). Thus, both the clustering result and the high variance of copy numbers indicate that the within-variance of Pietrain is higher than other breeds.

Whether domesticated or in the wild, most individuals were clustered along their region rather than their way of life. It implies that gene flow between domesticated pigs and wild boar is still occurring in some areas. Even with the separation between domesticated and wild, the impact of artificial selection on porcine CNV may not be large enough to surpass the impact of gene flow between domesticated and wild.

All the Woori-Heukdon (KWH) and Korean native pigs (KNP) were clustered together with Duroc. KWH was developed by crossbreeding of three generations starting from pure Duroc sow and KNP, also called Chookjin-Chamdon. The F1 hybrid sow was crossed with pure Duroc boar, and the F2 hybrid sow was crossed with Duroc boar. Because the breed development was a recent event finished in 2011, the inherited CNV of KWB has been changed a little.

The pairwise V_{ST} becomes smaller when V_S becomes larger. Variance of copy number was the largest in Pietrain among investigated breeds. Therefore, V_{ST} of pairs of Pietrain and other breeds had the smallest V_{ST}. In contrast, all pairs with Enshi black pig had the highest V_{ST}. Due to the fact that the distance between breeds in clustering on the heatmap was only measured with the mean value of pairwise V_{ST}, the clustering of breeds was not always concordant with their groups.

Copy number alteration of genes can make drastic change in phenotype by affecting on the expression and the structure of protein. Therefore, the copy number differentiated genes would be suggested as candidate regions of selection. I suggested how CNVs involved in the evolution of each population by considering environmental differences between respective population and functions of copy number differentiated genes.

Polycystic Kidney and Hepatic Disease 1-Like 1 (*PKHD1L1*) encodes a member of the polycystin protein family containing 11 transmembrane domains. *PKHD1L1* has been reported as a candidate gene for variation in pH of pork (Chung et al., 2015), which is related to meat color and water holding capacity. The average copy numbers of *PKHD1L1* were slightly lost in groups except for NEW, and they were slightly higher in the European than Asian population. This CNV would be a causative variation on the difference in meat color and water holding capacity between populations.

CLEC4E encodes C-type lectin domain family 4 member E protein. The protein, also called Mincle (Macrophage inducible C-type lectin), is an innate immune receptor on myeloid cells sensing pathogens (Patin et al., 2017). Since it was first described as a receptor for mycobacterial cell wall glycolipid and cord factor, the role of Mincle in innate immunity against mycobacterial infection has been investigated. Upregulation of Mincle expression in response to mycobacterial infection were observed in mice (Behler et al., 2012). When Mincle senses the motif of microbial signal, it induces pro-inflammatory responses. In addition to this fundamental role as a receptor, Mincle can act as an immune modulator in different models by either promoting anti-inflammatory cytokines expression or downregulating pro-inflammatory signaling pathways (Ostrop & Lang, 2017; Patin et al., 2017). Tuberculosis, mainly caused by mycobacterial infection, is a severe threat to pigs. Wild boar was suggested as a reservoir that maintains and spreads tuberculosis infection (Cowie et al., 2016). The copy numbers of *CLEC4E* were lost in domestic groups and NEW while neutral in EW and AW. The higher copy number of the *CLEC4E* in wild boars may be presented as evidence of adaptation to mycobacterial infection prevalent in the wild environment.

The average copy number of early endosome antigen 1 encoding gene, *EEA1* in every groups was more than 8.8 (Table 3.7). These abnormal copy numbers are most likely caused by minor variations in the reference genome. I demonstrated average copy numbers of five groups in genomic regions, including upstream, protein coding, and downstream region of *EEA1* in Figure 3.5. The average copy numbers in all groups peaked in two regions: 90227001 – 90240000 and 90244001 – 90250000. Furthermore, the homologous shape of the graphs among all groups also supported the possibility of minor deletion in the reference genome. *EEA1* consists of 5' upstream, 31 exons, 30 introns, and 3' downstream sequences, and the peak regions covered exons 16-21, 23, 24 and their adjacent introns. The previous gene reconstruction using additional alternate transcripts of pig individuals also improved a model of *EEA1* whose model was missed in Ensembl (Gilbert, 2019).

The *GVIN1*, interferon-induced very large GTPase 1, was upregulated in *PRRSV*-infected porcine alveolar macrophage (Chaudhari et al., 2021) while downregulated in lungs during bacterial respiratory infection (Mortensen et al., 2011). However, the biological mechanism of *GVIN1* expression against infection and the phenotypical effect of deletion in the porcine genome remain poorly understood.

Kojima and Degawa (Kojima & Degawa, 2014) demonstrated that *UGT2B31* expression was higher in male pigs when compared to female pigs and that testosterone treatment of castrated boars increased *UGT2B31* expression. Considering the above literature and gene expression network, Sahadevan et al. (Sahadevan et al., 2015) suggested that *UGT2B31* could play steroid metabolic roles in porcine androgen/androstenone metabolism. Sabmborski et al. (Samborski et al., 2013) also demonstrated a significant decrease in *UGT2B31* expression on day 14 of the pregnant pig. These previous studies continuously identified the role of *UGT2B31* in steroid hormone biosynthesis. The copy number of *UGT2B31* in EW and NEW groups were significantly gained. Moreover, *SC5D* is another gene involved in steroid biosynthesis, such that the expression of *SC5D* was upregulated in the pig ovary during the luteal phase (Park et al., 2022). The copy number of *SC5D* was significantly different in my rank- and variance-based test, and the average copy numbers were slightly higher in European pigs than in others. Therefore, these steroid syntheses related genes could be suggested as candidates which can make a difference in reproductive traits between porcine populations.

Cytochrome P450 (CYP) is a type of oxygenase. A previous study identified differences in the fatty acid composition of adipose tissues between Korean native and Yorkshire pigs (Choi et al., 2008). The significantly higher expression of CYP genes in Yorkshire was presented as the cause of lower arachidonic acid and higher cis-11,14,17-Eicosatrienoic acid, which are responsible for meat flavor. One of CYP isoforms *CYP2C36* was also suggested as copy number variable genes in my result. The mRNA levels of *CYP2C33*, *CYP2C49*, *CYP3A29,* and *CYP3A46* were reported as significantly different between Meishan and Landrace in 5-months pigs according to their sex (Kojima & Degawa, 2016). In addition to the different androgen levels, CNV could be suggested as another cause of differential expression of several CYPs. Because CYPs are also important in the drug metabolism of pigs, CNV of CYP should be considered when studying pigs as a model animal for drug metabolism.

There were NEW-specifically duplicated genes such as *EFHC1*, *ZWINT*, and *ROPN1L,* but little was revealed about their function in pig. Moreover, the number of NEW individuals here were only two, which was too few to suppose these genes play important role in evolution of NEW. In addition, previous studies were not enough to investigate the functional impact of copy number variation of like-genes such as *MARCKSL1, HSBP1L1,* and *NIF3L1* in the pig. Furthermore, the copy number of *MARCKS* and *HSBP1* were not significantly variable in both V_{ST} and the *One-way ANOVA* test. *MYO1H* had not been reported yet about their phenotype and genomic variation in *Sus scrofa*.

This chapter is under review in *Scientific Data*

as a partial fulfillment of Jisung Jang's Ph.D program,

Chapter 4. Chromosome-level genome assembly of Korean native cattle and pangenome graph of 14 bos taurus assemblies

4.1. Abstract

This study presents the first chromosome-level genome assembly of Hanwoo, an indigenous Korean breed of Bos taurus taurus. This is the first genome assembly of Asian taurus breed. Also, we constructed a pangenome graph of 14 B. taurus genome assemblies. The contig N50 was over 22 Mb, the scaffold N50 was over 89 Mb and a genome completeness of 95.8%, as estimated by BUSCO using the mammalian set, indicated a high-quality assembly. 48.7% of the genome comprised various repetitive elements, including DNAs, tandem repeats, long interspersed nuclear elements, and simple repeats. A total of 27,314 protein-coding genes were identified, including 25,302 proteins with inferred gene names and 2,012 unknown proteins. The pangenome graph of 14 B. taurus autosomes revealed 528.47 Mb non-reference regions in total and 61.87 Mb Hanwoo-specific regions. Our Hanwoo assembly and pangenome graph provide valuable resources for studying B. taurus populations.

4.2. Background & Summary

Hanwoo is a native Korean taurine cattle breed with a 5000-year history as a draft animal for farming and transportation (Lee et al., 2014). In a short period, Hanwoo underwent significant changes in its demographic history and selection. During the Korean war (1950-1953), the number of Hanwoo dropped to about 390,000, but recovered to 1.02 million by the late 1950s. With the development of the South Korean economy and agricultural industry, Hanwoo transitioned from a draft to a meat production breed in the 1960s. Modern breeding programs, including performance tests, artificial insemination and genomic selection were initiated by the South Korean government in the 1980s. These programs have improved carcass weight and meat quality of Hanwoo by increasing intramuscular fat (marbling). As a result of continuous artificial selection, Hanwoo has gained unique features both in genome and traits.

This study presents a high-quality assembly of Hanwoo which is the first chromosome-level genome assembly of Asian Bos taurus taurus using a combination of PacBio Hifi, Isoform and Illumina RNA sequencing, with scaffold N50 length of 89 Mb. The completeness of the genome was confirmed by the BUSCO score of 95.8%. The top 31 scaffolds are all greater than 17 Mb in size with a total length of 2.69 Gb. 48.7% of the Hanwoo genome is composed of various repetitive elements. The genome was annotated to contain 27,314 protein-coding genes, including 25,302 proteins with inferred gene names and 2,012 unknown proteins.

I generated a pangenome graph of 14 high-quality *Bos taurus* autosomes including high-quality genome assemblies of Hanwoo, Hereford, Angus, Brown

Swiss, Highland, Holstein, Jersey, Original Braunvieh, Piedmontese, Simmental, Brahman, Nellore, N'Dama, and Ankole. I identified non-reference regions and breed-specific regions through pangenome graph. In Hanwoo, 528.47 Mb of total non-reference nodes and 61.87 Mb of Hanwoo-specific nodes were identified. This pangenome graph would be used to extract structural variations and make insightful observations among various populations of *Bos taurus*.

4.3. Materials and Methods

4.3.1. Sample collection and extraction of genomic DNA and RNA

The samples used in the study of Hanwoo genome included blood, sirloin, liver, and subcutaneous fat from a steer named "bull *2050*". The samples were collected from the Experimental farm of College of Agriculture and Life Sciences at Seoul National University, Pyeongchang-gun, Gangwon-do, Republic of South Korea (Figure 4.1) and were approved by the Seoul National University Institutional Animal Care and Use Committee (SNU-201129-1-1). It was castrated in 9.4 months of age, slaughtered and sampled in 32 months of age. All blood sampling was carried out by trained veterinarians, according to the approved institutional protocols. Genomic DNA were extracted from whole blood using Wizard Genomic DNA Purification kit following the manufacturer's protocol.

Sirloin, liver and subcutaneous fat tissues of Hanwoo bull *2050* were collected immediately after slaughter and frozen using liquid nitrogen and stored in a deep freezer until RNA extraction. RNA was isolated using the RNeasy kits (Qiagen, Valencia, CA) following the manufacturer's protocol.

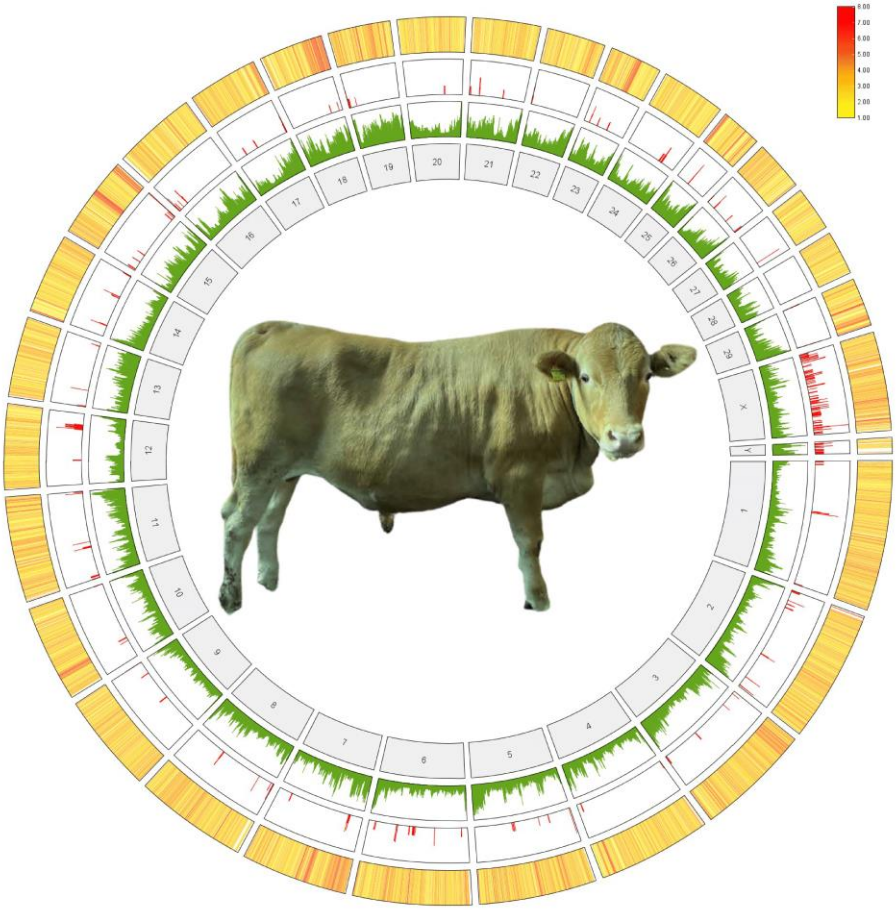

Figure 4.1. Circos plot denoting gene density, N ratio and GC content of Hanwoo genome assembly

4.3.2. DNA library Construction and sequencing

DNA sequencing libraries were prepared using SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences, California, USA) and libraries larger than 20kb were used for next steps. HiFi reads were sequenced using 2 SMRT cells of 8M Tray, Sequel II Sequencing Kit 2.0 in Pacific Biosciences (PacBio) Sequel IIe platform at NICEM in Seoul National University. Highly accurate consensus sequences were produced by PacBio *CCS* workflow (v 6.3.0), yielding a total of 3.5M reads and 67.5Gbp corresponding to a genomic coverage of ~24.8X (Table 4.1).

Table 4.1. Statistics of sequencing data

Platform	Tissue	Reads	Total bases (bp)	Average length (bp)	N50 length (bp)	SRA accession
PacBio	Blood	3,520,375	67,520,132,790	19180	20224	SRR23238456
RNA-seq	Liver	37986259	5773911368	76	76	SRR23238454
	Subcutaneous fat	37619668	5718189536	76	76	SRR23238453
	Sirloin	40572880	6167077760	76	76	SRR23238455
Iso-Seq	Sirloin	10,054,509	20,639,745,850	2,052	2,268	SRR23238452

4.3.3. RNA library Construction and sequencing

For RNA-seq, paired-end libraries with insert size of 75 bp were prepared with TruSeq Stranded mRNA Sample Preparation kit (Illumina, San Diego CA USA) from total messenger RNA (mRNA) of sirloin, liver and subcutaneous fat tissues of a Hanwoo bull 2050. RNA of the three tissues were sequenced separately using Illumina NextSeq 500 with following adapters; liver: D701, D506; sirloin: D701, D507; subcutaneous fat: D701, D508. 17.65 Gb of short paired-end RNA reads were sequenced using Illumina NextSeq 500 (Table 4.1).

For Iso-Seq, a total of 600 ng RNA from sirloin was used for full-length transcript sequencing with Pacbio Sequel system (Pacific Biosciences, CA, USA) according to the manufacturer's instructions. The Iso-Seq library was prepared according to the Isoform Sequencing (Iso-Seq) protocol using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module, PacBio SMRTbell Express Template Prep Kit 2.0 and ProNex® Size-Selective Purification System.

Total 10 μL library was prepared using PacBio SMRTbell Express Template Prep Kit 2.0. SMRTbell templates were annealed using Sequel Binding and Internal Ctrl Kit 3.0. The Sequel Sequencing Kit 3.0 and SMRT cells 1M v3 LR Tray was used for sequencing. SMRT cells (Pacific Biosciences) using 1200 min movies were captured for each SMRT cell using the PacBio Sequel System (Pacific Biosciences).

4.3.4. Genome size estimation and contig assembly

Hanwoo contigs were assembled using the HiFi consensus reads and validated following the VGP (Vertebrate Genomes Project) assembly pipeline (Lariviere et al., 2022). Adapter sequences of HiFi reads (5'–

ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT–3’) were removed by Cutadapt (v 4.0) (Martin, 2011). Counting k-mer and generating histogram of the k-mer count were performed on adapter trimmed sequences with *k*=21 by Meryl (v 1.3.0) (Rhie, 2020). Genome properties such as genome size, maximum read depth and transition parameter were inferred using GenomeScope (v 2.0) (Ranallo-Benavidez, Jaron, & Schatz, 2020) from 21-mer histogram generated by Meryl (v 1.3.0) (Rhie, 2020). Genome size of Hanwoo was estimated as 3.06 Gb based on the *k*-mer histogram (Figure 4.2). Trimmed reads were assembled to contig level using Hifiasm (v 0.16.1) (Cheng, Concepcion, Feng, Zhang, & Li, 2021), and the draft primary contig assembly consisted of 1311 contigs totaling 3.28 Gb with an N50 of 55.23 Mb (Table 4.2). Haplotypic duplication and low-coverage contigs of the draft contig assembly were removed using Purge_dups (v 1.2.5) (Guan et al., 2020) after self-alignment using Minimap2 (Li, 2018). The primary contig assembly after removing haplotypic duplication included 603 contigs, with a size of 3.11 Gb and a contig N50 of 58.14 Mb.

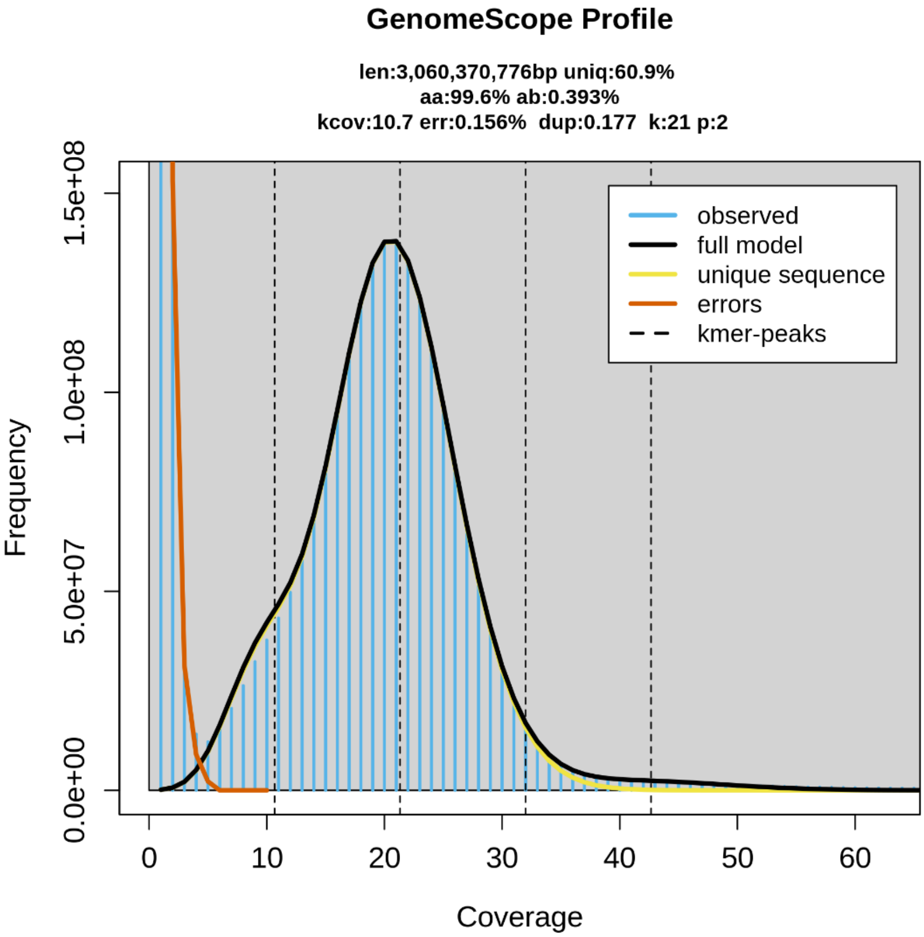

Figure 4.2. k-mer spectra and genome size estimation of Hanwoo by GenomeScope2

Table 4.2. Statistics of contig assembly before scaffolding

Statistics	**Draft primary contig assembly**	**Draft alternate contig assembly**	**Purged primary contig assembly**	**Purged alternate contig assembly**
Number of contigs	2342	10506	1053	1410
Largest contig	78136331	4548020	78136331	4260193
Total length	3469213782	2576051545	3141335384	327878398
N50	22442903	690552	24214384	443137
N75	5869989	285910	10061475	176266
L50	43	1048	36	163
L75	114	2462	86	465
GC (%)	44.34	43.16	43.58	51.61

4.3.5. Scaffolding and gap filling

The Hanwoo contigs after removing haplotypic duplication were scaffolded on autosome of *ARS-UCD1.3*, through reference-guided approach by RagTag (v 2.1.0) (Alonge et al., 2021). Because the Y chromosome is absent in *ARS-UCD1.3*, autosome and X chromosome of *ARS-UCD1.3,* and Y chromosome of *UOA_Angus_1* were used as reference genome for scaffolding. The reference-guided scaffolding using RagTag (v 2.1.0) (Alonge et al., 2021) consist of 'correct' and 'scaffold' steps. The 'correct' step identified and corrected potential misassembly based on alignment of contig assembly to the reference genome assembly. Part of contigs were broken at points of putative misassembly, and as a result, the number of contigs increased to 1915. In the 'scaffold' step, these RagTag 'corrected' contigs were aligned to the reference genome consist of autosome and X chromosome of *ARS-UCD1.3,* and Y chromosome of *UOA_Angus_1*. As a result, there were 1598 scaffolds including 31 chromosome-level scaffolds and 1567 unplaced scaffolds.

HiFi reads used in the Hanwoo assembly were aligned using Minimap2 (Li, 2018) to perform gap filling of the chromosome-level Hanwoo genome assembly using TGS-GapCloser (v 1.0.1) (Xu et al., 2020). The final 31 chromosome-level scaffolds had a total size of 2.69 Gb, which was similar to chromosome size of *ARS-UCD 1.3*. (Table 4.3, Table 4.4). These 31 chromosome-level scaffolds composed 86.66% of the assembly, with the remaining 414.6 Mb still unanchored and requiring further investigation. Further analysis including annotation and pangenome were performed on the chromosome-level scaffolds.

Table 4.3. Hanwoo genome assembly statistics

Assembly statistics	**Value**
Genome size (bp)	3139631388
Number of scaffolds	1599
Number of chromosome-scale scaffolds	31
N50 of scaffolds (bp)	88220521
L50 of scaffolds	14
Chromosome-scale scaffolds (bp)	2720843998
GC content of the genome (%)	43.58
QV score	64.15
Error rate	3.84E-07
BUSCO analysis	
Library	mammalia_odb10
Complete	8835 (95.7%)
Complete and single copy	8648 (93.7%)
Complete and duplicated	187 (2.0%)
Fragmented	108 (1.2%)
Missing	283 (3.1%)

Table 4.4. Length of Chromosome-level scaffolds

Chromosome	**Length**	**% of assembly**
1	159930546	5.88
2	141937731	5.22
3	122773356	4.51
4	124015044	4.56
5	122387257	4.50
6	121546567	4.47
7	112384917	4.13
8	115759820	4.25
9	107050878	3.93
10	105696925	3.88
11	108870483	4.00
12	90136002	3.31
13	86409007	3.18
14	84332089	3.10
15	86161463	3.17
16	89552414	3.29
17	74527950	2.74
18	69453907	2.55
19	66082172	2.43
20	72354257	2.66
21	79220027	2.91
22	61635694	2.27
23	54472963	2.00
24	63946808	2.35
25	43196348	1.59
26	53975766	1.98
27	47270444	1.74
28	46468219	1.71
29	52549481	1.93
X	139059706	5.11
Y	17685757	0.65
Total	2720843998	100.00

Circos plot denoting gene density, N ratio and GC content was generated with the advanced circos function from Java-based tool TBtools (Chen et al., 2020) (Figure 4.1). The gene density (number of genes), N ratio (%) and GC content (%) was calculated for every 10,000 bp increment of the genome and was visualized in a heatmap format for gene density and histogram format for N ratio and GC content using BIN size 100,000.

4.3.6. Masking repetitive sequences

Repetitive sequences in the gap-filled Hanwoo assembly were soft-masked using RepeatMasker (v 4.1.5) (N. Chen, 2004) with a known library (cow) in Dfam (v 3.7) and RepBase (v 10/26/2018) using RMBlast. Repetitive elements predicted by RepeatMasker contained 1.31 Gb of sequences, accounting for 48.7 % of the genome, including 27.6%, 11.6%, 4.9%, 2.1% and 1.5% for LINEs, SINEs, LTR elements, DNA elements, and satellite repeats, respectively (Table 4.5).

Table 4.5. Statistics of repetitive elements

Class	Subclass	Number	Total length (bp)	% of genome
SINEs:		2094753	313564131	11.6
	MIRs	400500	57658494	2.13
LINEs:		1330892	748009858	27.66
	LINE1	593655	344177051	12.73
	LINE2	255372	65668302	2.43
	L3/CR1	34977	7228441	0.27
	RTE	445684	330755520	12.23
LTR elements:		427451	135515056	5.01
	ERVL	77626	30593842	1.13
	ERVL-MaLRs	124708	40777241	1.51
	ERV_classI	86198	37895001	1.4
	ERV_classII	120569	21931489	0.81
DNA elements:		299386	59032346	2.18
	hAT-Charlie	168309	31215648	1.15
	TcMar-Tigger	46961	12256395	0.45
Unclassified:		3226	495315	0.02
Total interspersed repeats:			1256616706	46.48
Small RNA:		255446	43273484	1.6
Satellites:		8408	40214282	1.49
Simple repeats:		535375	22402828	0.83
Low complexity:		81799	4048187	0.15
Total bases masked:			1324164230	48.97

4.3.7. Genome annotation

Illumina RNA-seq reads were trimmed to remove adapter sequences and low-quality bases using Trimmomatic (v 0.39) (Bolger, Lohse, & Usadel, 2014). The BRAKER3 (v 3.0.3) pipeline (Gabriel et al., 2023) was used for structural annotation of Hanwoo genome. The pipeline utilized three sources of extrinsic evidence; short-read RNA-seq (Illumina), protein sequences of Vertebrata in OrthoDB (v 11) (Kuznetsov et al., 2022) in addition to protein sequence of ARS-UCD1.3 to train Augustus (v 3.5.0) (Stanke et al., 2006) for gene prediction. Non-coding genes were predicted from tRNAscan-SE (v 2.0.12) (Chan, Lin, Mak, & Lowe, 2021) including Infernal (Nawrocki & Eddy, 2013).

The predicted gene sets were searched in 2 public functional databases, Swiss-Prot of UniProtKB (Bairoch & Apweiler, 2000) and Pfam (v 35.0) database (Mistry et al., 2021) to identify the potential function with BLASTP (v 2.13.0+) (Camacho et al., 2009) and functional domains with InterProScan (v 5.57) (Jones et al., 2014). I used scripts included in MAKER (v 3.01.03) (Campbell, Holt, Moore, & Yandell, 2014) to integrate functional annotations into structural annotations. The genome annotation was evaluated using BUSCO (v 5.3.2) (Simão et al., 2015) analysis with the conserved core set of mammalian genes, yielding a completeness score of 87.9%. A total of 27,314 protein-coding genes were identified, including 25,302 genes with inferred names and 2,012 unknown proteins.

4.3.8. Assessment of the chromosome-level genome assembly

N50, L50 and lengths of the chromosome-level Hanwoo genome assembly was calculated by QUAST (v 5.0.2) (Gurevich, Saveliev, Vyahhi, & Tesler, 2013). Single copy gene completeness was assessed with BUSCO (v 5.3.2) (Simão et al., 2015), using the metaeuk backend against 'mammalia_odb10'. Quality values (QV) was calculated with Merqury (v 1.3) (Rhie, Walenz, Koren, & Phillippy, 2020), with *k*-mer databases (*k*=21) constructed by Meryl (v 1.3) (Rhie, 2020).

4.3.9. Pangenome graph construction

The pangenome graph of 14 Bos taurus genomes, including the Hanwoo assembly, was generated using the Minigraph-Cactus Pangenome Pipeline (v 2.5.1) (Armstrong et al., 2020). 14 assemblies were collected with the Hereford assembly, *ARS-UCD1.3* (Rosen et al., 2020), as the reference genome. 8 haplotype-resolved assemblies of Angus (*UOA_Angus_1*, GCF_002263795.3), Brahman (*UOA_Brahman_1*) (Koren et al., 2018), Simmental (*ARS-Simm1.0*) (Heaton et al., 2021), Scottish Highland bull (*ARS_UNL_Btau-highland_paternal_1.0_alt*, GCA_009493655.1) (Rice et al., 2020), N'Dama (*ROSLIN_BTT_NDA1*), Ankole (*ROSLIN_BTI_ANK1*) (Talenti et al., 2022), Jersey (*ARS-LIC_NZ_Jersey*, GCA_021234555.1), Holstein Friesian (*ARS-LIC_NZ_Holstein-Friesian_1*, GCA_021347905.1) were obtained from NCBI. Original Braunvieh, Nellore, Brown Swiss, and Piedmontese were collected from the public database (https://doi.org/10.5281/ZENODO.5906579) and scaffolded and merged by RagTag (Alonge et al., 2021) following the protocol of previous article (Leonard et al., 2022). The repeat sequences in the genomes of Original Braunvieh, Nellore, Brown Swiss, Piedmontese and Highland were soft-masked for by RepeatMasker (v 4.1.5) (N. Chen, 2004) using same parameters and repeat databases with Hanwoo. Because one sex chromosome was missing in haplotype-resolved genomes produced by trio-binning assembly, only autosomes were included in my pangenome graph.

The Minigraph-Cactus Pangenome Pipeline consisted of four steps: constructing the Minigraph GFA, mapping the genomes back to the Minigraph, creating the

Cactus alignment and creating the VG indexes. The Minigraph graph was created using ARS-UCD1.3 as the reference genome, and the other 13 genomes were iteratively added. Base-level alignments of the genomes were added to the graph using Cactus (Armstrong et al., 2020After embedding the haplotypes into the graph, Cactus alignment were performed, resulting in variation graph (VG) and hierarchical alignment (HAL). The HAL file was converted to packed graph (PG) and chopped into 32 base pairs using 'hal2vg' to describe it as nodes and edges. The HAL file was also converted to multiple alignment format (MAF) and synteny identified from MAF using 'maf2synteny' (Kolmogorov et al., 2018). The synteny diagram was generated using genomic coordinates of syntenic regions for three cattle breeds as input parameters for Python package pyGenomeViz (v 0.3.2, https://github.com/moshi4/pyGenomeViz). The pangenome graph was visualized by using 'vg view' (Garrison et al., 2018).

4.3.10. Non-reference nodes in pangenome graph

The multiple whole-genome alignments generated by CACTUS (Armstrong et al., 2020) were transformed into the Packed Graph (PG) format by chopping into 32 base pairs using 'hal2vg' with the options '--chop 32' and '--noAncestors' (Hickey, Paten, Earl, Zerbino, & Haussler, 2013). The reference nodes and non-reference nodes were separated using scripts from the Github repository (https://github.com/evotools/CattleGraphGenomePaper/tree/master/detectSequences/nf-GraphSeq) following previous research (Talenti et al., 2022). After excluding nodes flanking with gaps in 1kb, the counts and lengths of non-reference and breed-specific nodes were calculated (Table 4.6). Non-reference region and Hanwoo-specific regions longer and equal to 10kb are marked in Hanwoo autosome using KaryoploteR (Gel & Serra, 2017) (Figure 4.3). The Hanwoo-specific regions are included in non-reference region. Most of them were distributed in telomeric region. It suggested that larger genome and specific region of Hanwoo are result from expansion in repeat-rich telomeric region. In addition, HiFi-based assemblies generally have higher telomeric completeness than Oxford nanopore- or CLR-based assemblies (Leonard et al., 2023). The uniqueness of origin and evolution history also supported the larger and disctinct genome of Hanwoo compared to European taurine. Mitochondrial DNA haplogroup of Hanwoo is P, which is common in European aurochs but has not been detected in modern cattle in Europe(Achilli et al., 2008). The haplogroup P mtDNA in Hanwoo suggested the possibility of a minor and local event of domestication or introgression of Asian aurochs (Mannen et al., 2020; Noda et al., 2018). Furthermore, intensive inbreeding and small effective population size of Hanwoo might facilitate fixation of these distinctive regions in Hanwoo genome (Li & Kim, 2015).

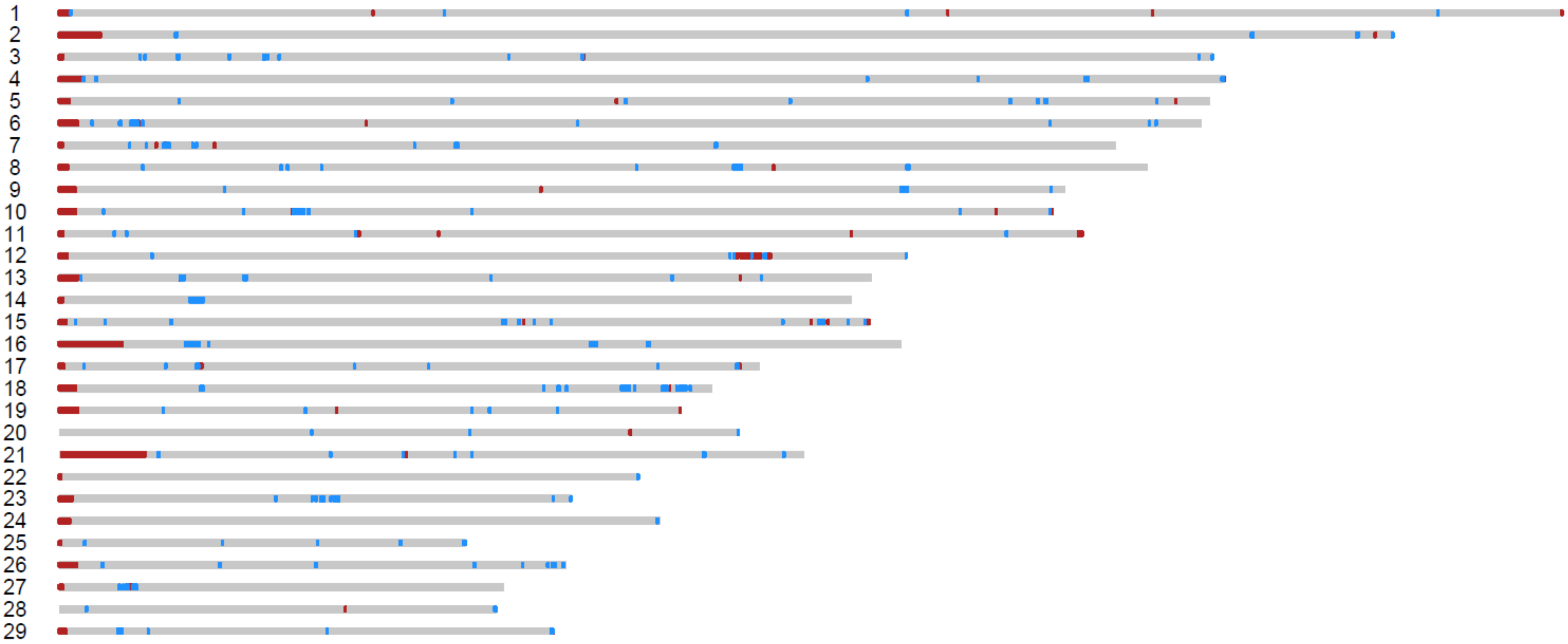

Figure 4.3. Non-reference region and specific region in Hanwoo autosome.

Non-reference regions and Hanwoo-specific regions larger than or equal to 10kb are visualized on Hanwoo autosomes. The Hanwoo-specific regions are marked in red, while the non-reference regions shared by other Bos taurus assemblies, excluding the Hanwoo-specific regions, are marked in blue.

Table 4.6. Sequence contribution of 14 bos taurus autosomes in the pangenome

Breed	**Non-reference nodes**		**Specific nodes**		**Total length (autosome)**
	nodes	**bp**	**nodes**	**bp**	bp
Hanwoo	5644829	83917034	622052	61869953	2538711408
Angus	4876028	40793146	331609	23589072	2468157877
Brown Swiss	5135844	25626114	364958	8631263	2497220059
Highland	4917533	32014564	383674	14515221	2483452092
Holstein	5046695	31095517	434031	16204587	2468170459
Jersey	5050922	27795391	402709	11095169	2473656513
Orininal Braunvieh	5135877	27234395	361737	10537892	2503654516
Piedmontese	5128788	28520430	389915	11411557	2500499917
Simmental	5266669	40554393	527318	20773580	2494093306
Brahman	11480493	46633118	2650315	20140251	2478073158
Nellore	12648594	45129061	3423881	19092260	2502536439
N'Dama	7225426	54175845	1375922	35064951	2504036093
Ankole	8960222	44980693	1959559	23916971	2485084605
Hereford					2489385779

4.4. Data Records

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JARDUZ000000000.

The transcriptomic Illumina sequencing data of subcutaneous fat, liver and sirloin were deposited in the SRA at NCBI SRR23238453, SRR23238454 and SRR23238455, respectively.

The transcriptomic PacBio sequencing data of sirloin were deposited in the SRA at NCBI SRR23238452.

The final chromosome assembly was deposited in NCBI BioProject PRJNA927262.

The genome annotation file (Jang, 2023) and the pangenome graph (Jang, 2023) are available in Figshare.

4.5. Technical Validation

RNA degradation and contamination were monitored on Agilent RNA ScreenTape. The purity of RNA samples was checked using the NanoPhotometer spectrophotometer (IMPLEN, CA, USA). The integrity of RNA was assessed using the RNA ScreenTape of the Agilent 2200 TapeStation System (Agilent Technologies, CA, USA). Only RNAs with an OD260/280 ratio of 2.0-2.2, an OD260/230 ratio of 1.8-2.1, and a RIN value of ≥9.0 were considered qualified for use. RNA concentration was measured using Quant-iT™ RiboGreen™ RNA Assay Kit in Victor Nivo (PerkinElmer, Waltham, MA, USA).

The completeness of the Hanwoo genome assembly was evaluated using BUSCO (Simão et al., 2015) with the mammalian data set “mammalia_odb10.” The evaluation found 95.7% (8835) of the core mammalian genes were present in the genome, including 93.7% single-copy, 2.0% duplicated, 1.2% fragmental, and 3.1% missing genes from the mammalian data set (Table 4.3). The k-mer databases (k=21) constructed using HiFi reads by Meryl (Rhie, 2020), and the overall assembly quality was assessed using the k-mer databases using Merqury(Rhie et al., 2020). The assembly showed high quality values (QV > 64) with an error rate of 3.84×10^(-7) (Table 5.3). The GC content of Hanwoo (43.58%) was similar to that of ARS-UCD1.3 (41.56%). These assessment results confirmed the completeness of Hanwoo genome assembly (Table 4.3).

To validate the Hanwoo genome assembly and Hanwoo-specific regions, Illumina short reads from additional Hanwoo individuals were aligned to the Hanwoo genome assembly and Hanwoo-specific region separately using ‘vg giraffe’ (Garrison et al., 2018) (Table 5.7). Whole-genome sequence reads from three Hanwoo individuals and cDNA reads of four Hanwoo individuals were mapped to Hanwoo genome and Hanwoo-specific regions, respectively using BWA-MEM2 (v 2.2.1) (Vasimuddin, Misra, Li, & Aluru, 2019). Mapping coverage of specific regions of Hanwoo was from 2.22 to 2.51%, slightly smaller than the proportion of specific regions in the Hanwoo genome, 3.50%. The higher mapping rate and coverage of DNA than cDNA to Hanwoo-specific regions suggeste that the larger portion of the specific regions consist of non-coding region such as repeats, rather than coding regions.

Table 4.7. Samples used in short read alignment on Hanwoo assembly

Sample accession	Run accession	Instrument model	Study accession	Gender type	Tissue	Mapping rate on Hanwoo genome (%)	Coverage on Hanwoo genome (%)	Mapping rate on Hanwoo-specific regions (%)	Coverage on Hanwoo-specific regions (%)	Mean Depth on Hanwoo genome
Genome										
SAMN02225729	SRR934400	Illumina HiSeq 2000	PRJNA210523	NA	Blood	99.70022824	93.4856499	40.5523161	2.492581515	3.901084002
SAMN02225732	SRR934401	Illumina HiSeq 2000	PRJNA210523	NA	Blood	99.69021678	94.27086924	41.72421157	2.505379223	3.820759449
SAMN02225731	SRR934404	Illumina HiSeq 2000	PRJNA210523	NA	Blood	99.71003797	79.25842694	40.40382882	2.225516894	1.834271754

Transcriptome (mRNA)										
SAMN01093740	SRR527009	Illumina HiSeq 2000	PRJNA171257	female	Subcutaneous fat	99.5880854	8.760859832	15.01287962	0.551694674	1.550169227
SAMN01093745	SRR527014	Illumina HiSeq 2000	PRJNA171257	castrated male	Intramuscular fat	99.62212433	6.634598858	14.58888742	0.450361186	1.370927687
SAMN01093761	SRR527030	Illumina HiSeq 2000	PRJNA171257	male	Muscle	99.60091206	5.774603323	19.89932617	0.397774445	1.322897901
SAMN01093743	SRR527012	Illumina HiSeq 2000	PRJNA171257	castrated male	Omental fat tissue	99.62203677	5.048693129	19.35225445	0.344376973	1.269874722

4.6. Usage Notes

One of the key benefits of pangenome graph is the visual identification of structural variation. I visualized pangenome graph near the copy number variable region (CNVR) which was identified in a previous study (Jisung Jang et al., 2021) (Figure 4.4). The pre-defined region is part of WD repeat domain 25 (WDR25), covering 65,311,801 to 65,315,200 in chromosome 21 of Hereford genome (ARS-UCD1.2). The copy numbers of CNVR were significantly different between Asian indicine and African taurine. The average copy number was 2.9 in Eurasian taurine, 3.3 in African taurine, 1.1 in African humped cattle, and 0.7 in Asian indicus, respectively. In a portion of the pangenome graph, I identified deletions of 15 nodes in Nellore, Brahman, and Ankole around the reference node '93074445' located at 65,315,183 of Hereford chromosome 21, which is included in the CNVR. This supports the population differentiated CNV of the previous study (Jisung Jang et al., 2021).

To identify insertion in Hanwoo genome, syntenic region adjacent to the specific region of Hanwoo were investigated. There were syntenic regions adjacent to a specific region of Hanwoo chromosome 18, from 14,513,559 to 14,592,390. Syntenic regions in upstream (chr18: 14,310,250- 14,513,324) and downstream (chr18: 14,592,584-14,973,625) of the Hanwoo-specific region were identified in all of the genomes included in my pangenome graph (Figure 4.5). This supports the insertion in Hanwoo genome.

These two examples about CNV and insertion serve as good examples of identifying structural variation using my pangenome graph.

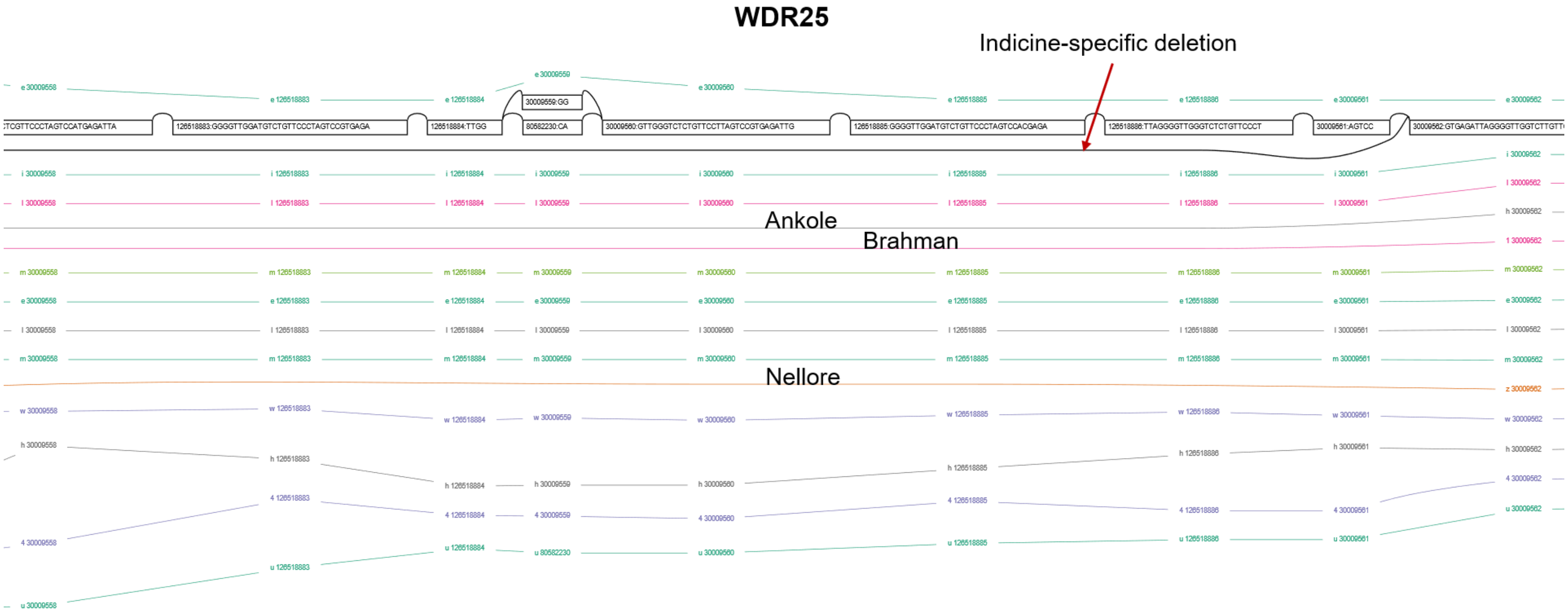

Figure 4.4. Copy number variation in WDR25 in pangenome graph

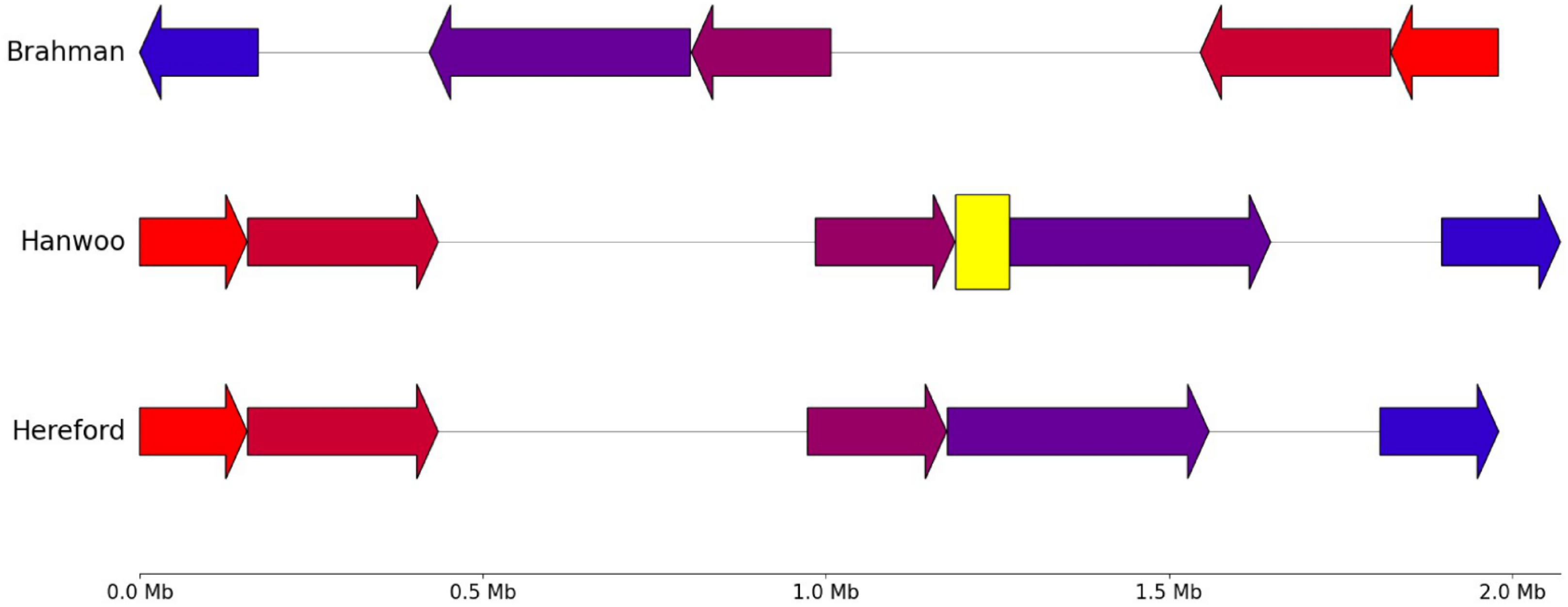

Figure 4.5. Insertion between syntenic region in Hanwoo chromosome 18, from 14,513,559 to 14,592,390

Chapter 5. General discussion

The population differentiation and characteristics of structural variation (SV) in livestock species plays a crucial role in understanding the evolution and disease susceptibility of these animals. This dissertation aimed to investigate the population genetics of SVs in cattle and swine, two important domesticated animals with complex evolutionary histories. The research presented in this dissertation utilized various bioinformatic approaches to analyze SV in three distinct chapters, focusing on CNV.

The first chapter of this dissertation provided a comprehensive literature review on structural variation and population genetics. It explored the fundamental concepts and methods used to study SVs in different populations of the same species. Understanding the nature and distribution of SVs in populations is essential for deciphering their evolutionary and functional implications.

Chapter 2 focused on population differentiated CNVs among Bos taurus, Bos indicus, and their African hybrids. The study revealed the impact of hybridization and selection on CNV diversity, shedding light on the genetic consequences of breed mixing and selective breeding practices. These findings contribute to my understanding of the genetic architecture underlying phenotypic variation and adaptation in cattle populations.

In Chapter 3, the CNV profiles of Eurasian wild boar and domesticated pig populations were compared. The analysis provided insights into the signatures of domestication and adaptation on CNV patterns in swine. Understanding the genetic changes associated with domestication is crucial for improving pig breeds and managing their genetic diversity.

In Chapter 4, the first chromosome-level genome assembly of Hanwoo, an indigenous Korean breed of Bos taurus taurus, was presented. This achievement marked the first genome assembly of an Asian taurus breed. Additionally, a pangenome graph of 14 B. taurus assemblies was constructed, revealing non-reference regions and Hanwoo-specific regions. The study identified structural variants and genetic elements that may be associated with phenotypic traits and adaptation. These genomic resources provide valuable tools for studying B. taurus populations and contribute to my understanding of the genetic diversity within this species.

Collectively, the chapters presented in this dissertation demonstrate the power and utility of population differentiation and characteristics of SVs for studying the evolution and disease of livestock species. By investigating CNV patterns, important insights into the genetic architecture, adaptation, and disease susceptibility of cattle and swine populations were gained. The findings of this dissertation contribute to the broader field of population genetics and provide valuable resources and insights for future research in livestock genomics and animal breeding.

References

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research, 21(6), 974-984.

Achilli, A., Bonfiglio, S., Olivieri, A., Malusa, A., Pala, M., Kashani, B. H., Perego, U. A., Ajmone-Marsan, P., Liotta, L., & Semino, O. (2009). The multifaceted origin of taurine cattle reflected by the mitochondrial genome. PloS one, 4(6), e5753.

Achilli, A., A. Olivieri, M. Pellecchia, C. Uboldi, L. Colli, N. Al-Zahery, M. Accetturo, M. Pala, B. H. Kashani and U. A. Perego (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. Current Biology 18(4): R157-R158.

Aguiar, T. S., Torrecilha, R. B. P., Milanesi, M., Utsunomiya, A. T. H., Trigo, B. B., Tijjani, A., Musa, H. H., Lopes, F. L., Ajmone-Marsan, P., & Carvalheiro, R. (2018). Association of copy number variation at intron 3 of HMGA2 with navel length in Bos indicus. Frontiers in genetics, 9, 627.

Ajmone-Marsan, P., Garcia, J. F., & Lenstra, J. A. (2010). On the origin of cattle: how aurochs became cattle and colonized the world. Evolutionary Anthropology: Issues, News, and Reviews, 19(4), 148-157.

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z., . . . Soyk, S. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. BioRxiv.

Andrews, S. (2017). FastQC: a quality control tool for high throughput sequence data. 2010.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., . . . Stiller, J. (2020). Progressive Cactus is a multiple-genome aligner for the thousandgenome era. Nature, 587(7833), 246-251.

Bahbahani, H., Afana, A., & Wragg, D. (2018). Genomic signatures of adaptive introgression and environmental adaptation in the Sheko cattle of southwest Ethiopia. PloS one, 13(8).

Bahbahani, H., Tijjani, A., Mukasa, C., Wragg, D., Almathen, F., Nash, O., Akpa, G. N., Mbole-Kariuki, M., Malla, S., & Woolhouse, M. (2017). Signatures of selection for environmental adaptation and zebu× taurine hybrid fitness in East African Shorthorn Zebu. Frontiers in genetics, 8, 68.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research, 28(1), 45-48.

Behler, F., Steinwede, K., Balboa, L., Ueberberg, B., Maus, R., Kirchhof, G., Yamasaki, S., Welte, T., & Maus, U. A. (2012). Role of Mincle in alveolar macrophage-dependent innate immunity against mycobacterial infections in mice. The Journal of Immunology, 189(6), 3121-3129.

Bera, A., Singh, S., Nagaraj, R., & Vaidya, T. (2003). Induction of autophagic cell death in Leishmania donovani by antimicrobial peptides. Molecular and biochemical parasitology, 127(1), 23-35.

Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., Song, J., Schnabel, R. D., Ventura, M., & Taylor, J. F. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. Genome research, 22(4), 778-790.

Bickhart, D. M., Xu, L., Hutchison, J. L., Cole, J. B., Null, D. J., Schroeder, S. G., Song, J., Garcia, J. F., Sonstegard, T. S., & Van Tassell, C. P. (2016). Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Research, 23(3), 253-262.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Boligon, A., De Vargas, L., Silveira, D., Roso, V., Campos, G., Vaz, R., & Souza, F. (2016). Genetic models for breed quality and navel development scores and its associations with growth traits in beef cattle. Tropical animal health and production, 48(8), 1679-1684.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC bioinformatics, 10(1), 1-9.

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Current protocols in bioinformatics, 48(1), 4.11. 11-14.11. 39.

Canal, L. B., Fontes, P. L., Sanford, C. D., Mercadante, V. R., DiLorenzo, N., Lamb, G. C., & Oosthuizen, N. (2020). Relationships between feed efficiency and puberty in Bos taurus and Bos indicus-influenced replacement beef heifers. Journal of Animal Science.

Chan, E. K., Nagaraj, S. H., & Reverter, A. (2010). The evolution of tropical adaptation: comparing taurine and zebu cattle. Animal Genetics, 41(5), 467-477.

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic acids research, 49(16), 9077-9096.

Chaudhari, J., Liew, C.-S., Riethoven, J.-J. M., Sillman, S., & Vu, H. L. (2021). Porcine Reproductive and Respiratory Syndrome Virus Infection Upregulates Negative Immune Regulators and T-Cell Exhaustion Markers. Journal of virology, 95(21), e01052-01021.

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. Molecular plant, 13(8), 1194-1202.

Chen, N. (2004). Using Repeat Masker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics, 5(1), 4.10. 11-14.10. 14.

Chen, S., Lin, B.-Z., Baig, M., Mitra, B., Lopes, R. J., Santos, A. M., Magee, D. A., Azevedo, M., Tarroso, P., & Sasazaki, S. (2010). Zebu cattle are an exclusive legacy of the South Asia Neolithic. Molecular biology and evolution, 27(1), 1-6.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature methods, 18(2), 170-175.

Choi, K.-M., Moon, J.-K., Choi, S.-H., Kim, K.-S., Choi, Y.-I., Kim, J.-J., & Lee, C.-K. (2008). Differential expression of cytochrome P450 genes regulate the level of adipose arachidonic acid in Sus Scrofa. Asian-Australasian Journal of Animal Sciences, 21(7), 967-971.

Chung, H., Lee, K., Jang, G., Choi, J., Hong, J., & Kim, T. (2015). A genome-wide analysis of the ultimate pH in swine. Genet Mol Res, 14(4), 15668-15682.

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., ... & Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. Nature, 581(7809), 444-451.

Conrad, D. F., & Hurles, M. E. (2007). The population genetics of structural variation. Nature genetics, 39(Suppl 7), S30-S36.

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

Consortium, G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. nature, 491(7422), 56.

Cowie, C. E., Hutchings, M. R., Barasona, J. A., Gortázar, C., Vicente, J., & White, P. C. (2016). Interactions between four species in a complex wildlife: livestock disease community: implications for Mycobacterium bovis maintenance and transmission. European journal of wildlife research, 62(1), 51-64.

Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J., Alcalá, A. M., Sonstegard, T. S., Hanotte, O., Götherström, A., Seabury, C. M., & Praharani, L. (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet, 10(3), e1004254.

Edea, Z., Bhuiyan, M., Dessie, T., Rothschild, M., Dadi, H., & Kim, K. (2015). Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. Animal, 9(2), 218-226.

Elsik, C. G., Tellam, R. L., & Worley, K. C. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science, 324(5926), 522-528.

Erixon P, Svennblad B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics Systematic Biology 52:665-673

Fang, M., Larson, G., Soares Ribeiro, H., Li, N., & Andersson, L. (2009). Contrasting mode of evolution at a coat color locus in wild and domestic pigs. PLoS genetics, 5(1), e1000341.

Frantz, L. A., Haile, J., Lin, A. T., Scheu, A., Geörg, C., Benecke, N., Alexander, M., Linderholm, A., Mullin, V. E., & Daly, K. G. (2019). Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. Proceedings of the National Academy of Sciences, 116(35), 17231-17238.

Frantz, L. A., Schraiber, J. G., Madsen, O., Megens, H.-J., Bosse, M., Paudel, Y., Semiadi, G., Meijaard, E., Li, N., & Crooijmans, R. P. (2013). Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome biology, 14(9), 1-12.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., . . . Lin, M. F. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology, 36(9), 875-879.

Gel, B. and E. Serra (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33(19): 3088-3090.

Ghoreishifar, S. M., Eriksson, S., Johansson, A. M., Khansefid, M., Moghaddaszadeh-Ahrabi, S., Parna, N., Davoudi, P., & Javanmard, A. (2020). Signatures of selection reveal candidate genes involved in economic traits and cold acclimation in five Swedish cattle breeds. Genetics Selection Evolution, 52(1), 1-15.

Gilbert, D. G. (2019). Genes of the pig, Sus scrofa, reconstructed with EvidentialGene. PeerJ, 7, e6374.

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., & Megens, H.-J. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature, 491(7424), 393-398.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36(9), 2896-2898.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Hanotte, O., Bradley, D. G., Ochieng, J. W., Verjee, Y., Hill, E. W., & Rege, J. E. O. (2002). African pastoralism: genetic imprints of origins and migrations. Science, 296(5566), 336-339.

Hayes, B. J., MacLeod, I. M., Daetwyler, H. D., Bowman, P. J., Chamberlain, A. J., Vander Jagt, C., Capitan, A., Pausch, H., Stothard, P., & Liao, X. (2014). Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. Proceedings of the 10th world congress of genetics applied to livestock production,

Heaton, M. P., Smith, T. P., Bickhart, D. M., Vander Ley, B. L., Kuehn, L. A., Oppenheimer, J., . . . McClure, J. C. (2021). A reference genome assembly of Simmental cattle, Bos taurus taurus. Journal of Heredity, 112(2), 184-191.

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics, 29(10), 1341-1342.

Hou, Y., Bickhart, D. M., Chung, H., Hutchison, J. L., Norman, H. D., Connor, E. E., & Liu, G. E. (2012). Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. Functional & integrative genomics, 12(4), 717-723.

Hu, Y., Xia, H., Li, M., Xu, C., Ye, X., Su, R., Zhang, M., Nash, O., Sonstegard, T. S., & Yang, L. (2020). Comparative analyses of copy number variations between Bos taurus and Bos indicus. BMC genomics, 21(1), 1-11.

Hu, Z.-L., Park, C. A., & Reecy, J. M. (2019). Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic acids research, 47(D1), D701-D710.

Jang, J. (2023). Bos taurus pangenome graph. figshare https://doi.org/10.6084/m9.figshare.21273609

Jang, J. (2023). Hanwoo Genome Assembly (Bos taurus). figshare https://doi.org/10.6084/m9.figshare.22086665

Jang, J., Jung, J., Lee, Y.H., Lee, S., Baik, M., Kim, H. (2023). Bos taurus breed Hanwoo isolate HWB-2050, whole genome shotgun sequencing project. GenBank https://identifiers.org/ncbi/insdc:JARDUZ000000000

Jang, J., Terefe, E., Kim, K., Lee, Y. H., Belay, G., Tijjani, A., . . . Kim, H. (2021). Population differentiated copy number variation of Bos taurus, Bos indicus and their African hybrids. BMC genomics, 22(1), 1-11.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . . Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236-1240.

Kampinga, H. H., & Craig, E. A. (2010). The HSP70 chaperone machinery: J proteins as drivers of functional specificity. Nature reviews Molecular cell biology, 11(8), 579-592.

Kasarapu, P., Porto-Neto, L. R., Fortes, M. R., Lehnert, S. A., Mudadu, M. A., Coutinho, L., Regitano, L., George, A., & Reverter, A. (2017). The Bos taurus–Bos indicus balance in fertility and milk related genes. PloS one, 12(8), e0181930.

Keel, B. N., Lindholm-Perry, A. K., & Snelling, W. M. (2016). Evolutionary and functional features of copy number variation in the cattle genome. Frontiers in genetics, 7, 207.

Kim, J., Hanotte, O., Mwai, O. A., Dessie, T., Bashir, S., Diallo, B., Agaba, M., Kim, K., Kwak, W., & Sung, S. (2017). The genome landscape of indigenous African cattle. Genome biology, 18(1), 1-14.

Kim, J.-H., Hu, H.-J., Yim, S.-H., Bae, J. S., Kim, S.-Y., & Chung, Y.-J. (2012). CNVRuler: a copy number variation-based case–control association analysis tool. Bioinformatics, 28(13), 1790-1792.

Kim, K., Kwon, T., Dessie, T., Yoo, D., Mwai, O. A., Jang, J., Sung, S., Lee, S., Salim, B., Jung, J., Jeong, H., Tarekegn, G. M., Tijjani, A., Lim, D., Cho, S., Oh, S. J., Lee, H.-K., Kim, J., Jeong, C., . . . Kim, H. (2020). The mosaic genome of indigenous African cattle as a unique genetic resource for African pastoralism. Nature Genetics. https://doi.org/10.1038/s41588-020-0694-2

Kojima, M., & Degawa, M. (2014). Sex differences in the constitutive gene expression of sulfotransferases and UDP-glucuronosyltransferases in the pig liver: androgen-mediated regulation. Drug Metabolism and Pharmacokinetics, 29(2), 192-197.

Kojima, M., & Degawa, M. (2016). Sex differences in constitutive mRNA levels of CYP2B22, CYP2C33, CYP2C49, CYP3A22, CYP3A29 and CYP3A46 in the pig liver: Comparison between Meishan and Landrace pigs. Drug Metabolism and Pharmacokinetics, 31(3), 185-192.

Kolde, R. (2012). Pheatmap: pretty heatmaps. R package version, 1(2), 726.

Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., Dunn, M., Yang, F., . . . Thybert, D. (2018). Chromosome assembly of large and complex genomes using multiple references. Genome research, 28(11), 1720-1732.

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., . . . Phillippy, A. M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. Nature biotechnology, 36(12), 1174-1182.

Kulkarni, M. M., Barbi, J., McMaster, W. R., Gallo, R. L., Satoskar, A. R., & McGwire, B. S. (2011). Mammalian antimicrobial peptide influences control of cutaneous Leishmania infection. Cellular microbiology, 13(6), 913-923.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2022). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic acids research.

Lariviere, D., Ostrovsky, A., Gallardo, C., Syme, A., Abueg, L., Pickett, B., . . . Sozzoni, M. (2022). VGP assembly pipeline.

Larson G et al. (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication Science 307:1618-1621

Larson, G., Liu, R., Zhao, X., Yuan, J., Fuller, D., Barton, L., Dobney, K., Fan, Q., Gu, Z., & Liu, X.-H. (2010). Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. Proceedings of the National Academy of Sciences, 107(17), 7686-7691.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome biology, 15(6), 1-19.

Lee, K., Nguyen, D. T., Choi, M., Cha, S.-Y., Kim, J.-H., Dadi, H., Seo, H. G., Seo, K., Chun, T., & Park, C. (2013). Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. BMC genomics, 14(1), 596.

Lee, S.-H., Park, B.-H., Sharma, A., Dang, C.-G., Lee, S.-S., Choi, T.-J., . . . Kim, S.-D. (2014). Hanwoo cattle: origin, domestication, breeding strategies and genomic selection. Journal of animal science and technology, 56(1), 1-8.

Leonard, A. S., Crysnanto, D., Fang, Z.-H., Heaton, M. P., Vander Ley, B. L., Herrera, C., . . . Smith, T. P. (2022). Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. Nature communications, 13(1), 1-13.

Leonard, A. S., D. Crysnanto, X. M. Mapel, M. Bhati and H. Pausch (2023). Graph construction method impacts variation representation and analyses in a bovine super-pangenome. Genome Biology 24(1): 124.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: visualization of intersecting sets. IEEE transactions on visualization and computer graphics, 20(12), 1983-1992.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100.

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

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

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., Wang, T., Yeung, C. K., Chen, L., & Ma, J. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nature genetics, 45(12), 1431-1438.

Li, Y. and J.-J. Kim (2015). Effective population size and signatures of selection using bovine 50K SNP chips in Korean native cattle (Hanwoo). Evolutionary Bioinformatics 11: EBO. S24359.

Li, Z., Gilbert, J. A., Zhang, Y., Zhang, M., Qiu, Q., Ramanujan, K., Shavlakadze, T., Eash, J. K., Scaramozza, A., & Goddeeris, M. M. (2012). An HMGA2-IGF2BP2 axis regulates myoblast proliferation and myogenesis. Developmental cell, 23(6), 1176-1188.

Liu, G. E., & Bickhart, D. M. (2012). Copy number variation in the cattle genome. Functional & integrative genomics, 12(4), 609-624.

Liu, G. E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J., & Eichler, E. E. (2009). Analysis of recent segmental duplications in the bovine genome. BMC genomics, 10(1), 571.

Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M., & Cunningham, P. (1994). Evidence for two independent domestications of cattle. Proceedings of the National Academy of Sciences, 91(7), 2757-2761.

Low, W. Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D. M., Rosen, B. D., Kronenberg, Z. N., Kingan, S. B., & Tseng, E. (2020). Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nature communications, 11(1), 1-14.

Ma, Y.-L., Wen, Y.-F., Cao, X.-K., Cheng, J., Huang, Y.-Z., Ma, Y., Hu, L.-Y., Lei, C.-Z., Qi, X.-L., & Cao, H. (2019). Copy number variation (CNV) in the IGF1R gene across four cattle breeds and its association with economic traits. Archives animal breeding, 62(1), 171-179.

Magee, D. A., MacHugh, D. E., & Edwards, C. J. (2014). Interrogation of modern and ancient genomes reveals the complex domestic history of cattle. Animal Frontiers, 4(3), 7-22.

Mallikarjunappa, S., Sargolzaei, M., Brito, L. F., Meade, K. G., Karrow, N., & Pant, S. (2018). Uncovering quantitative trait loci associated with resistance to Mycobacterium avium ssp. paratuberculosis infection in Holstein cattle using a high-density single nucleotide polymorphism panel. Journal of dairy science, 101(8), 7280-7286.

Mannen, H., T. Yonezawa, K. Murata, A. Noda, F. Kawaguchi, S. Sasazaki, A. Olivieri, A. Achilli and A. Torroni (2020). Cattle mitogenome variation reveals a post-glacial expansion of haplogroup P and an early incorporation into northeast Asian domestic herds. Scientific Reports 10(1): 20842

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), 10-12.

Megens, H.-J., Crooijmans, R. P., San Cristobal, M., Hui, X., Li, N., & Groenen, M. A. (2008). Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genetics Selection Evolution, 40(1), 1-26.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic acids research, 47(D1), D419-D426.

Mielczarek, M., Frąszczak, M., Nicolazzi, E., Williams, J., & Szyda, J. (2018). Landscape of copy number variations in Bos taurus: individual–and inter-breed variability. BMC genomics, 19(1), 410.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., . . . Richardson, L. J. (2021). Pfam: The protein families database in 2021. Nucleic acids research, 49(D1), D412-D419.

Moioli, B., D'Andrea, S., De Grossi, L., Sezzi, E., De Sanctis, B., Catillo, G., Steri, R., Valentini, A., & Pilla, F. (2016). Genomic scan for identifying candidate genes for paratuberculosis resistance in sheep. Animal Production Science, 56(7), 1046-1055.

Mortensen, S., Skovgaard, K., Hedegaard, J., Bendixen, C., & Heegaard, P. M. (2011). Transcriptional profiling at different sites in lungs of pigs during acute bacterial respiratory infection. Innate Immunity, 17(1), 41-53.

Nakamura, Y., Kanemarum, K., & Fukami, K. (2013). Physiological functions of phospholipase Cδ1 and phospholipase Cδ3. Advances in Biological Regulation, 53(3), 356-362.

NCBI Sequence Read Archive, (2023). https://identifiers.org/ncbi/insdc.sra:SRP419181.

Nicholas, T. J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E. E., & Akey, J. M. (2009). The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome research, 19(3), 491-499.

Niimura, Y. (2012). Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. Current genomics, 13(2), 103-114.

Noda, A., R. Yonesaka, S. Sasazaki and H. Mannen (2018). The mtDNA haplogroup P of modern Asian cattle: A genetic legacy of Asian aurochs? PLoS One 13(1): e0190937.

Oldham J (1972) Epidemic Diarrhea—How it all began Pig Farming:72-73

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., & Ako-Adjei, D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), D733-D745.

Ostrop, J., & Lang, R. (2017). Contact, collaboration, and conflict: signal integration of Syk-coupled C-type lectin receptors. The Journal of Immunology, 198(4), 1403-1414.

Park, Y., Park, Y.-B., Lim, S.-W., Lim, B., & Kim, J.-M. (2022). Time Series Ovarian Transcriptome Analyses of the Porcine Estrous Cycle Reveals Gene Expression Changes during Steroid Metabolism and Corpus Luteum Development. Animals, 12(3), 376.

Patin, E. C., Orr, S. J., & Schaible, U. E. (2017). Macrophage inducible C-type lectin as a multifunctional player in immunity. Frontiers in immunology, 8, 861.

Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L. A., Bosse, M., Crooijmans, R. P., & Groenen, M. A. (2015). Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. BMC genomics, 16(1), 1-14.

Pierce, M. D., Dzama, K., & Muchadeyi, F. C. (2018). Genetic diversity of seven cattle breeds inferred using copy number variations. Frontiers in genetics, 9, 163.

Rabelo, R. E., Silva, L. A. F. d., Brito, L. A. B., Moura, M. I. d., Silva, O. C. d., Carvalho, V. S. d., & Franco, L. G. (2008). Epidemiological aspects of surgical diseases of the genital tract in a population of 12,320 breeding bulls (1982-2007) in the state of Goias, Brazil.

Rambaut A (2012) FigTree version 1.4. 0 Available at ht tp://tree bio ed ac uk/software/figtree

Ranallo-Benavidez, T., Jaron, K., & Schatz, M. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. In: Nature Publishing Group.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., & Chen, W. (2006). Global variation in copy number in the human genome. nature, 444(7118), 444-454.

Reed, D. R., & Knaapila, A. (2010). Genetics of taste and smell: poisons and pleasures. In Progress in molecular biology and translational science (Vol. 94, pp. 213-240). Elsevier.

Revilla, M., Puig-Oliveras, A., Castello, A., Crespo-Piazuelo, D., Paludo, E., Fernandez, A. I., Ballester, M., & Folch, J. M. (2017). A global analysis of CNVs in swine using whole genome sequence data and association analysis with fatty acid composition and growth traits. PLoS One, 12(5), e0177014.

Rhie, A. (2020). Meryl. GitHub repository: GitHub.

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome biology, 21(1), 1-27.

Rice, E. S., Koren, S., Rhie, A., Heaton, M. P., Kalbfleisch, T. S., Hardy, T., . . . Ley, B. V. (2020). Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. Gigascience, 9(4), giaa029.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models Bioinformatics 19:1572-1574

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., Rowan, T. N., Low, W. Y., Zimin, A., & Couldrey, C. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience, 9(3), giaa021.

Rubin, C.-J., Megens, H.-J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., & Jørgensen, C. B. (2012). Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences, 109(48), 19529-19536.

Sadkowski, T., Jank, M., Zwierzchowski, L., Siadkowska, E., Oprządek, J., & Motyl, T. (2008). Gene expression profiling in skeletal muscle of Holstein-Friesian bulls with single-nucleotide polymorphism in the myostatin gene 5'-flanking region. Journal of Applied Genetics, 49(3), 237-250.

Sahadevan, S., Tholen, E., Große-Brinkhaus, C., Schellander, K., Tesfaye, D., Hofmann-Apitius, M., Cinar, M. U., Gunawan, A., Hölker, M., & Neuhoff, C. (2015). Identification of gene co-expression clusters in liver tissues from multiple porcine populations with high and low backfat androstenone phenotype. BMC genetics, 16(1), 1-18.

Sainz, R., Cruz, G., Mendes, E., Magnabosco, C., Farjalla, Y., Araujo, F., Gomes, R., & Leme, P. (2013). Performance, efficiency and estimated maintenance energy requirements of Bos taurus and Bos indicus cattle. In Energy and protein metabolism and nutrition in sustainable animal production (pp. 69-70). Springer.

Samborski, A., Graf, A., Krebs, S., Kessler, B., & Bauersachs, S. (2013). Deep sequencing of the porcine endometrial transcriptome on day 14 of pregnancy. Biology of reproduction, 88(4), 84, 81-13.

Schiavo, G., Dolezal, M., Scotti, E., Bertolini, F., Calò, D., Galimberti, G., Russo, V., & Fontanesi, L. (2014). Copy number variants in Italian Large White pigs detected using high-density single nucleotide polymorphisms and their association with back fat thickness. Animal genetics, 45(5), 745-749.

Schlattl, A., Anders, S., Waszak, S. M., Huber, W., & Korbel, J. O. (2011). Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome research, 21(12), 2004-2013.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212.

Sjödin, P., & Jakobsson, M. (2012). Population genetic nature of copy number variation. In Genomic Structural Variants (pp. 209-223). Springer.

Spehr, M., & Munger, S. D. (2009). Olfactory receptors: G protein-coupled receptors and beyond. Journal of neurochemistry, 109(6), 1570-1583.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research, 34(suppl_2), W435-W439.

Stock, F., & Gifford-Gonzalez, D. (2013). Genetics and African cattle domestication. African Archaeological Review, 30(1), 51-72.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., & Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. Science, 330(6004), 641-646.

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., & Jones, P. G. (2007). A single IGF1 allele is a major determinant of small size in dogs. Science, 316(5821), 112-115.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics, 22(12), 1540-1542.

Talenti, A., Powell, J., Hemmink, J. D., Cook, E. A., Wragg, D., Jayaraman, S., . . . Agusi, E. (2022). A cattle graph genome incorporating global breed diversity. Nature communications, 13(1), 1-14.

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics, 31(12), 2032-2034.

Taye, M., Lee, W., Caetano-Anolles, K., Dessie, T., Cho, S., Jong Oh, S., Lee, H.-K., & Kim, H. (2018). Exploring the genomes of East African Indicine cattle breeds reveals signature of selection for tropical environmental adaptation traits. Cogent Food & Agriculture, 4(1), 1552552.

Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J. R., Sung, W. W., Pereira, S. L., Whitney, J., Chan, A. J., & Pellecchia, G. (2018). A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. The American Journal of Human Genetics, 102(1), 142-155.

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using QQ and manhattan plots. Biorxiv, 005165.

Upadhyay, M., Bortoluzzi, C., Barbato, M., Ajmone-Marsan, P., Colli, L., Ginja, C., Sonstegard, T. S., Bosse, M., Lenstra, J. A., & Groenen, M. A. (2019). Deciphering the patterns of genetic admixture and diversity in southern European cattle using genome-wide SNPs. Evolutionary applications, 12(5), 951-963.

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. Paper presented at the 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).

Vigne, J.-D. (2011). The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. Comptes rendus biologies, 334(3), 171-181.

Wang, Y., Reverter, A., Kemp, D., McWilliam, S., Ingham, A., Davis, C., Moore, R., & Lehnert, S. (2007). Gene expression profiling of Hereford Shorthorn cattle following challenge with Boophilus microplus tick larvae. Australian Journal of Experimental Agriculture, 47(12), 1397-1407.

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., Chow, W., Eory, L., Finlayson, H. A., & Flicek, P. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience, 9(6), giaa051.

White, S. (2011). From globalized pig breeds to capitalist pigs: a study in animal cultures and evolutionary history. Environmental History, 16(1), 94-120.

Wilkinson, S., Lu, Z. H., Megens, H.-J., Archibald, A. L., Haley, C., Jackson, I. J., Groenen, M. A., Crooijmans, R. P., Ogden, R., & Wiener, P. (2013). Signatures of diversifying selection in European pig breeds. PLoS genetics, 9(4), e1003453.

Xu, J., Fu, Y., Hu, Y., Yin, L., Tang, Z., Yin, D., Zhu, M., Yu, M., Li, X., & Zhou, Y. (2020). Whole genome variants across 57 pig breeds enable comprehensive identification of genetic signatures that underlie breed features. Journal of Animal Science and Biotechnology, 11(1), 1-16.

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., . . . Deng, L. (2020). TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience, 9(9), giaa094.

Zeder MA (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact Proceedings of the national Academy of Sciences 105:11597-11604

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. Annual review of genomics and human genetics, 10, 451-481.

Zheng, X., Zhao, P., Yang, K., Ning, C., Wang, H., Zhou, L., & Liu, J. (2020). CNV analysis of Meishan pig by next-generation sequencing and effects of AHR gene CNV on pig reproductive traits. Journal of Animal Science and Biotechnology, 11(1), 1-11.

국문초록

소와 돼지의 구조 변이의 특성 및 집단 간 차이 연구

장지성

협동과정 생물정보학전공

서울대학교 대학원 자연과학대학

구조 변이(structural variation, SV)는 1 kb보다 긴 DNA 영역의 변화를 포함하는 유전체 변이의 한 종류이다. 구조 변이는 유전자 발현, 기능에 영향을 미치며 다양한 형질과 질병과 관련되어 있으며, 진화의 역사 추정을 위한 단서이다. 본 연구에서는 복잡한 진화 역사를 가진 두 가지 중요한 가축인 소와 돼지의 구조 변이의 집단 유전학을 연구하였다. 유전차 상의 다양한 구조 변이 중에서, 특히 구간의 결실 또는 중복을 포함하는 구조 변이의 한 형태인 복제 수 변이(copy number variation, CNV)에 초점을 맞춘 3개의 주제들을 연구하기 위해 다양한 유전체학적, 생물정보학적 방법을 활용하였다.

제1장에서는 구조변이와 구조변이의 집단 유전학적 특성 및 분석 방법 등 본 논문에 포함된 기본 지식과 연구 동향을 정리하였다.

제2장에서는 Bos taurus, Bos indicus 및 그들의 교잡으로 형성된 아프리카 소들 간의 차별화된 복제 수 변이를 조사하여 교잡과 선택이 CNV 다양성에 미치는 영향을 밝혔다.

제3장에서는 유라시아 멧돼지와 가축화된 돼지 집단 간의 복제 수 변이를 비교하여 가축화와 적응에 따른 CNV 패턴의 특징을 발견하였다.

제4장에서는 한우의 염색체 수준의 고품질 genome assembly와 14개 *Bos taurus* 유전체들의 pangenome graph를 제시하였다. 이 연구에서 형질과 적응과 관련될 수 있는 한우 특이적 영역과 구조 변이를 확인하였다.

본 논문은 소와 돼지의 진화와 질병을 연구하기 위한 구조 변이의 집단 간 차이와 특성을 연구하여, 진화적 관점의 해석을 제공하였으며, 이는 향후 연구를 위한 귀중한 자료와 통찰을 제공하였다.

주요어: 유전체, 구조 변이, 복제 수 변이, 진화

학번: 2016-28977