Genome-wide copy number variation in sheep: detection and utility as a genetic marker for quantitative traits, with reference to gastrointestinal nematodiasis

Thesis presented in partial fulfilment of
the requirements for the degree of

Doctor of Philosophy
in
Animal science

At Massey University, Palmerston North,
New Zealand

Juncong Yan

2018

Abstract

Gastrointestinal nematodes are perhaps the most important parasites of domestic sheep world-wide. Genetic selection for nematode resistance in domestic sheep is being promoted in many countries including New Zealand. There are several strategies to identify genetic markers associated with quantitative traits. Single nucleotide polymorphism (SNP)-based strategies have been widely used in animal breeding. However, SNP cannot explain all the genetic variation for a particular trait. A new kind of variation, copy number variation (CNV) has been identified as contributing to genetic variation in production and disease traits. Compared with other domestic animals, CNV in sheep is poorly investigated. The primary objective of this thesis was to explore the utility of genome-wide CNV as a genetic marker for the analysis of quantitative traits in sheep. Five different studies were undertaken to fulfill the objective. The first two studies used 50 K SNP BeadChip genotype data and next generation sequencing (NGS) data to detect CNV. Extensive CNV differences were evident between breeds as well as detection algorithms. NGS-based detection resulted in better CNV resolution than that by SNP. Subsequently, a genome-wide association study (with a small sample size) using CNV detected from a high density (HD) SNP genotype data identified four CNV regions to be significantly associated with a couple of traits pertaining to gastrointestinal nematodiasis in Romney sheep, while no significant SNP associations were found. Somatic mosaicism of CNV, influenced by age (high in foetuses, compared to adults), individuals, detection algorithm and type of tissue analysed, was also evident in separate study. The final study detected CNV differences and SNP based selection signatures in two Romney lines selected for gastrointestinal nematode resistance or resilience. Several significant SNPs and line-specific CNV regions were identified. However, only one SNP overlapped to a CNV region, indicating that SNP-based selection signatures and CNV could represent different aspects of sheep immunogenetics. Overall, CNV could be a potential

genetic marker, albeit with methods for detection and validation needing to be refined. The conclusions from this thesis expand our understanding of CNV in sheep and its potential application prospects for genetic breeding of sheep in the future.

Acknowledgements

First and foremost, I would like to acknowledge my supervisors, Dr Rao Dukkipati, Prof Hugh Blair and Associate Prof Patrick Biggs. Rao, you like a father, endlessly supported and encouraged me during my PhD study. You had no complaints about my poor English and tolerated that for four years. Your passion, patience and meticulousness for science let me know what a scientist should be. Hugh, you like a grandfather, did your best to provide funds for my study. I cannot count how many times you have helped me. Your academic brilliance illuminated my academic path like a beacon and you have been my model in heart. Patrick, you opened the gate of bioinformatics for me and show me a total new interesting world.

Furthermore, I would like to acknowledge Prof Dorian Garrick, Dr Keren Dittmer, Dr Sarah Pain, Dr Andrew Greer, Mr Joseph Hamie for original data support and Dr Kristene Gedye, Rosemary Heathcott for technical support. I would also like to acknowledge the overseas laboratory group, Key Laboratory of Genetics Breeding and Reproduction of Grass Feeding Livestock; Key Laboratory of Animal Biotechnology of Xinjiang, Xinjiang Academy of Animal Science, in China. The data you supported is very important for my research. I also thank my all colleagues in room 3.06. Your support and harmonious environment in office has been crucial for my successful completion. Your help let me quickly adapt to New Zealand's life.

I would like to acknowledge research funding from different sources within Massey University as well as the Massey-Lincoln and Agricultural Industry Trust, for this project. Besides, I very much appreciate the financial support from the Massey University Doctoral scholarship, which really helped me a lot and let me focus on study. I also appreciate the financial support from the IVABS postgraduate fund which helped me to attend my first international academic conference in Australia.

I would like to acknowledge NeSi's (New Zealand eScience Infrastructure) support for high performance computing for the NGS data analysis.

Finally, I would like to thank my family. Dad, Mum and my brother, your encouragement is the power for me to climb the mountain of academic success. Your never-ending support has helped me succeed. Thank you for always believing in me.

Preface

I have undertaken this thesis in the form of publishable experimental chapters using a format of thesis by publication. The current status and publication outlet are described in the following list.

Chapter 1: Literature review

Chapter 2: Genome-wide detection of autosomal copy number variants in several sheep breeds using Illumina OvineSNP50 BeadChips.

Juncong Yan, Hugh T. Blair, Mingjun Liu, Wenrong Li, Sangang He, Lei Chen, Keren E. Dittmer, Dorian J. Garrick, Patrick J. Biggs, Venkata S.R. Dukkipati*
Published in Small Ruminant Research, 2017, 155: 24-32.
(doi:10.1016/j.smallrumres.2017.08.022)

All molecular work, data analysis, interpretation of results and manuscript write-up were completed by Juncong Yan. The original SNP data was provided by Mingjun Liu, Wenrong Li, Sangang He, Lei Chen, Keren E. Dittmer and Dorian J. Garrick. The manuscript was checked by supervisors, Venkata S.R. Dukkipati, Hugh T. Blair and Patrick J. Biggs.

Chapter 3: Detection of copy number variation in sheep by whole genome sequencing

Juncong Yan, Hugh T. Blair, Keren E. Dittmer, Patrick J. Biggs, Venkata S.R. Dukkipati
To be submitted to BMC Genomics

All molecular work, data analysis, interpretation of results and manuscript write-up were completed by Juncong Yan. The original NGS data was provided by Keren E. Dittmer. The manuscript was checked by supervisors, Venkata S.R. Dukkipati, Hugh T. Blair and Patrick J. Biggs.

Chapter 4: Genome-wide association study in sheep selectively bred for resistance or resilience to gastrointestinal nematodes

Juncong Yan, Hugh T. Blair, Andrew Greer, Joseph Hamie, Patrick Biggs, Venkata S.R. Dukkipati

To be submitted to Journal of Animal Breeding and Genetics

All molecular work, data analysis, interpretation of results and manuscript write-up were completed by Juncong Yan. The original phenotype data was provided by Andrew Greer and Joseph Hamie. Patrick J. Biggs provided bioinformatics support for gene annotation. The manuscript was checked by supervisors, Venkata S.R. Dukkipati, Hugh T. Blair and Patrick J. Biggs.

Chapter 5: Somatic mosaicism of copy number variation in sheep using Ovine Infinium® HD SNP BeadChip

Juncong Yan, Hugh T. Blair, Patrick J. Biggs, Sarah J. Pain, Venkata S.R. Dukkipati

To be submitted to PLOS ONE

All molecular work, data analysis, interpretation of results and manuscript write-up were completed by Juncong Yan. The original tissue samples were provided by Sarah J. Pain. Patrick J. Biggs provided bioinformatics support for gene annotation. The manuscript was checked by supervisors, Venkata S.R. Dukkipati, Hugh T. Blair and Patrick J. Biggs.

Chapter 6: Detection of copy number variation and genome-wide positive selection signatures using Ovine Infinium® HD SNP BeadChip in two Romney lines, selected for resistance or resilience to gastrointestinal nematodes

Juncong Yan, Hugh T. Blair, Andrew Greer, Joseph Hamie, Patrick J. Biggs, Venkata S.R. Dukkipati

To be submitted to BMC Genomics

Part of this chapter was presented as a poster at the 22nd Association for the Advancement of Animal Breeding and Genetics conference held in Townsville, QLD, Australia, 2-5 Jul 2017.

Juncong Yan, Venkata S.R. Dukkipati, Hugh T. Blair, Patrick Biggs, Joseph Hamie and Andrew Greer (2017). A genome-wide scan of positive selection signature using Ovine Infinium® HD SNP BeadChip in two Romney lines, selected for resistance or resilience to nematodes. In Proceedings of the Association for the Advancement of Animal Breeding and Genetics Vol. 22 (pp.1).

All molecular work, data analysis, interpretation of results and manuscript write-up were completed by Juncong Yan. The original phenotype data was provided by Andrew Greer and Joseph Hamie. The manuscript was checked by supervisors, Venkata S.R. Dukkipati, Hugh T. Blair and Patrick J. Biggs.

Chapter 7: General discussion

Table of Contents

List of Figures

List of Tables

Common abbreviations

aCGH	array comparative genomic hybridization
AFLP	amplified fragment length polymorphism
AMD	age-related macular degeneration
AS	de novo assembly of a genome
BAF	B allele frequency
BLUP	best linear unbiased prediction
BP	biological process
CC	cellular component
CFH	complement factor H gene
CIITA	class II Major Histocompatibility Complex transactivator
CN-LOH	mosaic copy neutral loss of heterozygosity
CNV	copy number variation
CNVR	copy number variation region
CPU	central processing unit
dH_2O	deionised distilled water
DLRS	derivative log ratio spread
DNA	deoxyribonucleic acid
dNTP	deoxy-ribonucleoside triphosphate
dsDNA	double-stranded DNA
EHH	extended Haplotype Homozygosity
ELISA	enzyme-Linked ImmunoSorbent Assay
EMP2	empirical p-value, corrected for all tests
EPG	eggs per gram
FDR	false discovery rate
FEC	faecal egg count
FLK	an extension of Lewontin and Krakauer (LK) test, based on population's kinship (F) matrix
Fst	fixation index
GO	gene ontology
GWAS	genome-wide association study
hapFLK	haplotype structure accounted FLK
HGP	human Genome Project
HIV	human immunodeficiency virus
HMMs	hidden Markov models
IBD	identity by descent
iHH	integrated allele-specific EHH
iHS	integrated haplotype Score
IQRs	inter-quartile range
ISGC	International Sheep Genomics Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	linkage disequilibrium

LRR	log R ratio
LW	live weights
MF	molecular function
MHC II	major histocompatibility complex II
MZ	monozygotic twins
NAHR	non-allelic homologous recombinations
NeSi	New Zealand eScience infrastructure
NGS	next generation sequencing
PCA	principal components analysis
PCR	polymerase chain reaction
PEM	paired-end mapping
qPCR	quantitative polymerase chain reaction
Q-Q	quantile-quantile
QTL	quantitative trait loci
RAM	random-access memory
RAPD	random Amplified Polymorphic DNA
RD	read depth
REHH	relative EHH
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
Rsb	across Population EHH
SLE	systemic lupus erythematosus
SM	somatic mosaicism
SNP	single nucleotide polymorphism
SR	split read
SRFA	selective restriction fragment amplification
SSRs	simple sequence repeats
SVS	Golden Helix SNP & Variation Suite
TMB	tetramethyl benzidine
VNTR	variable number of tandem repeats
WF	Wave factor
XP-EHH	across Population EHH
ZHp	Z-transformed Heterozygosity Value

Chapter 1

Literature review

1.1 Introduction

The New Zealand sheep industry is highly regarded around the world and creates about $2.56 billion (2017-2018) in export earnings each year (Beef+Lamb NZ 2017-18). Improvement of sheep has always been one of the most important research topics in New Zealand. Gastrointestinal nematodes are perhaps the most important parasites of domestic sheep world-wide. Anthelmintic chemotherapy aimed at curtailing nematode infections has become increasingly expensive and led to drug resistance (Roos 1997). Also, increasing concerns regarding drug residues in food have led a growing number of producers adopting husbandry programs that are less dependent on drugs. For these reasons, genetic selection for parasite resistance in domestic sheep is being promoted in many countries including New Zealand.

There are several strategies to identify genetic markers associated with quantitative traits. Genome-wide association study (GWAS) is one of them. Various phenotypes (e.g. resistance/susceptibility to nematode parasites) have been found to be associated with several genes rather than just one. The old research method for single candidate genes does not work well. In this situation, GWAS provides a path to deal with this issue. Simply, GWAS is a kind of statistical method to find the association between genes and traits at genome level. The first successful GWAS was undertaken by Klein et al. (2005), who found a single nucleotide polymorphism (SNP) variant in the complement factor *H* gene to strongly associate with age-related macular degeneration in humans. Nowadays, GWAS based on SNPs is being widely used in animal breeding as well as genetic epidemiology.

However, SNP cannot explain all the genetic variation for a particular trait. A new kind of variation, copy number variation (CNV) has been identified as contributing genetic variation in production and disease traits. CNVs are defined as segments of DNA (larger than 1 kb) displaying copy number differences such as gains (insertions or duplications) or losses (deletions or null genotypes) (Feuk et al. 2006; Scherer et al. 2007). It is estimated that thousands of genes (about 12% of the human genome) are variable in copy number and are likely to be responsible for a significant proportion of normal phenotypic variation (Carter 2007). Therefore, it is necessary to expand SNP based GWAS to CNV based GWAS. During the past decade, several studies in humans and animals have identified CNV polymorphisms and their association with quantitative as well as complex disease traits. Significant CNV relationships identified in livestock include: 1) duplication of agouti signaling protein gene, that results in white coat color in sheep (Norris and Whan 2008) and goat (Fontanesi et al. 2009) and 2) duplication of CIITA, a trans-activator of MHC II, associated with nematode resistance in Angus cattle (Liu et al. 2011). Prior to the commencement of this study, only two studies (Fontanesi et al. 2011; Liu et al. 2013) investigated CNV in sheep.

This PhD research primarily focused on the detection of CNV in sheep, based on SNP chip and next-generation sequencing data, and the utility of CNV as a genetic marker for quantitative traits, with reference to gastrointestinal nematodiasis. CNV found to be associated with nemoatodiasis can be employed by breeders as reference genetic markers to select nematode resistant sheep. Besides, they could also provide a clue to reveal potential mechanism of parasite related immue response, in further studies of molecular and cell biology.

1.2 Genetic markers

1.2.1 Introduction

Genetic markers are variations in either DNA sequence or single base pair changes with known

physical locations on chromosomes, which are popular research areas of livestock breeding. By studying the associations between genetic markers and traits, it is possible to find new pathways to treat disease and improve animal production. Until now, three generations of genetic markers, more than 30 kinds of markers, have been utilized (Maheswaran 2004). In this review, six kinds of typical markers are introduced below.

1.2.2 The history of genetic markers

1.2.2.1 The first generation genetic markers

1.2.2.1.1 Restriction fragment length polymorphism (RFLP)

DNA restriction enzymes can recognize specific sequences in DNA and cut DNA within those recognition sequences into smaller pieces, which can be displayed by electrophoresis in agarose gels and separated by their differences in length. Different individuals, within a species, will show different electrophoretic patterns on gels because of mutations wihtin the restriction enzyme cutting site (Botstein et al. 1980). In majority of the cases, RFLP patterns are not readily visible on the gel, instead the fragments from the gel are blotted on to a nylon membrane (Southern blotting) and then hybridized to probe that spans the polymorphic restriction site. The first publication on RFLP was in 1974, when Grodzicker et al. (1974) used this technology to study linkage of temperature-sensitive mutations in adenoviruses. The first RFLP map of an entire human genome was made by Donis-Keller et al. (1987). After that, RFLP became one of the most popular genetic markers and achieved widespread use in many areas, such as agriculture (Beckmann and Soller 1983; Sreenan et al. 1997; Tanksley et al. 1989; Velasquez and Gepts 1994). However, due to design of RFLP probes, RFLPs are expensive and labour intensive (Powell et al. 1996).

1.2.2.1.2 Variable Number of Tandem Repeats (VNTR)

VNTR or minisatellite is a location in a genome where short nucleotide sequence (ranging in

length from 10–60 base pairs) are typically repeated 5-50 times (MeSH concept M0027878). The first time VNTR came into public view was in 1987 in Britain as bio-evidence of homicide in a court case which was provided by Jeffreys et al. (1985a, 1985b), who found DNA "fingerprints" of human and developed VNTR as a genetic marker. The detection of VNTR is similar to the detection of RFLP. The difference between VNTR and RFLP is that the length of VNTR is dependent on the number of repeats rather than the mutation of the restriction enzyme cutting site as in RFLP. VNTR has been widely used in animal and plant genetic breeding (Hillel et al. 1990; Nybom 1994). However, VNTR has a few natural drawbacks. Like RFLPs, they are expensive and labour intense. Also, the restriction enzyme must be chosen carefully so that it will not break integrity of repeat zone.

1.2.2.2 The second generation genetic markers

1.2.2.2.1 Random Amplified Polymorphic DNA (RAPD)

In order to overcome the disadvantages of RFLP, Williams et al. (1990) and John Welsh et al. (1990) invented a new genetic marker in 1990 to get a genetic map called random amplified polymorphic DNA, using arbitrarily primed polymerase chain reaction (AP-PCR). The working theory of this technology is to use hundreds of random primers to amplify whole genomes and get the DNA pattern which will be used to study differences among individuals. Due to the ease of detection and speed of use of RAPD, this technology has been frequently used in plants (Reiter et al. 1992; Tingey et al. 1994). However, RAPD can be influenced by many factors and the lack of reproducibility requires standardization of conditions (Lowe et al. 1996). Besides, RAPD has dominant expression which cannot be used to find heterozygotes (Primrose and Twyman 2009).

1.2.2.2.2 Amplified fragment length polymorphism (AFLP)

Amplified fragment length polymorphism (AFLP), also called selective restriction fragment

amplification (SRFA), was invented by Vos et al. (1995) in 1993. AFLP is a diagnostic fingerprinting technology to detect genomic restriction fragments very much like RFLP (Voss et al. 1995). However, AFLP detects the presence or absence of restriction fragments rather than length differences. Because it does not use reference sequence, it has had widespread use in plant research (Primrose and Twyman 2009).

1.2.2.2.3 Microsatellites (SSRs)

Microsatellites or simple sequence repeats (SSRs) or short tandem repeats (STRs) can be defined as relatively short runs of tandemly repeated DNA with repeat lengths of 6 bp or less (Z Wang et al. 1994). The discovery of SSRs can be traced to 1974, when Skinner et al. (1974) found repeated DNA sequence (TAGG)n/(ATCC)n in crab (Pagurus policaris). After that, in 1981, Miesfeld et al. (1981) found STRs in human genomes. One year later, SSRs were discovered widely in eukaryotic genomes from yeast to human by Hamada et al. (1982). Tautz & Renz (1984) found these sequences just existed in eukaryotic genomes and named it as simple sequence repeat (SSR) in 1984. In 1989, microsatellites were firstly used by Litt et al. (1989). in the cardiac muscle actin gene. Because SSRs are widely distributed throughout eukaryotic genomes and have great diversity, it has been used widely in many fields, such as building genetic maps, diversity analysis and genetic breeding in the past 30 years.

1.2.2.3 The third generation genetic markers

1.2.2.3.1 Single nucleotide polymorphisms (SNP)

Single nucleotide polymorphisms (SNPs) are single base-pair variations of sequence in genomic DNA at which different sequence alternatives (alleles), insertions and deletions exist in the population (Cho et al. 1999; Primrose and Twyman 2009). There are more than 3 million SNPs estimated in the human genome (Cooper et al. 1985; Hofker et al. 1986), in other words, in each 1000 base pairs 1 SNP can be found. Until 2001, there were 1.42 million single

nucleotide polymorphisms found througout the human genome, most discovered by the SNP Consortium (TSC) and the public Human Genome Project (HGP). The average density of SNPs was 1 every 1.9 kilobases (Lander et al. 2001; Sachidanandam et al. 2001; Venter et al. 2001). Six years later, with the help of the international Hapmap project, the number of human SNPs rose to 3.1 million (Frazer et al. 2007). There are some advantages with SNPs. Firstly, SNPs use the change of code of DNA sequence as a genetic marker rather than the change of length of DNA sequences, such as with RFLP. Secondly, analysis of SNP uses array technology instead of gel electrophoresis, which dramatically increased data size output from genetic markers. Based on these reasons, the SNP has become one of the most important and popular genetic markers.

1.2.3 Summary

Several kinds of genetic markers have been found. RFLPS, VNTRs and microsatellites were quite popular during late 20th century. However, from the beginning of this century, SNPs have become the most popular genetic marker, benefited from automation and standardization of SNP detection and analysis. With the development of technology, more and more new genetic markers will be detected that will expand our view into new areas of genetics.

1.3 Copy number variation (CNV)

1.3.1 Introduction

Copy number variation is defined as stretches of DNA larger than 1 kb that display copy number differences as gains (insertions or duplications) or losses (deletions or null genotypes) in normal populations (Feuk et al. 2006; Scherer et al. 2007) as illustrated in Figure 1.1. The research about CNV can be traced back to 1936 when Calvin Bridges, who found an association between the BAR “gene”, a part of the chromosome, with the size of eye in *Drosophila melanogaster* (Bridges 1936). However, without molecular biology, most research could only

be done by observing the structure of the chromosome under a microscope.

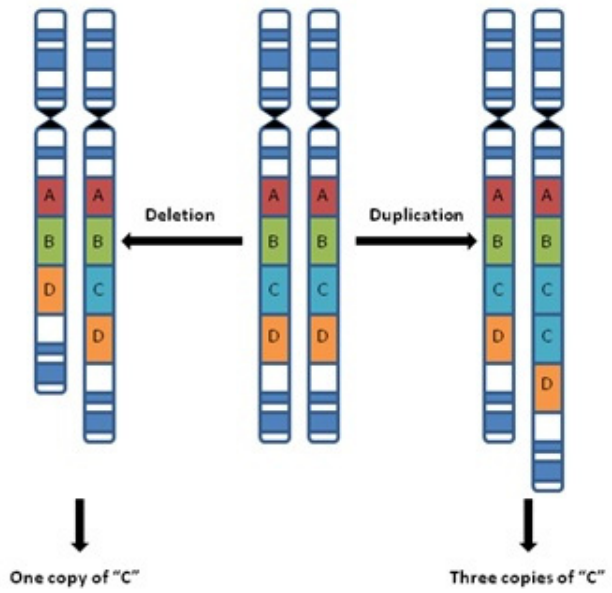


Figure 1.1 An explanation of copy number variation.

A normal pair of chromosomes each have sections A-B-C-D. However, the loss of section C from one of the chromosomes results in an abnormal chromosome with only sections A-B-D (left pair in the Figure); On the other hand, the gain of an extra copy of section C on one of the chromosomes results in an abnormal chromosome with sections A-B-C-C-D (right pair in the Figure) (Room 2018)

With the development of molecular biology, the molecular structure of copy number variation became known (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005). The first-generation CNV map of the human genome was constructed by Redon et al. (2006) by using Affymetrix GeneChip Human Mapping 500K early access array and they found 1447 CNVs. One year later, the impact of CNVs on gene expression variation was first detected in human lymphoblastoid cell lines by Stranger et al. (2007) who found that the complexity of functionally relevant genetic variation ranged from single nucleotides to megabases. After that, research on genome wide CNV association with traits became a hot topic in genetic research.

1.3.2 Function of CNV

It has been proposed that the effects of CNV differences on phenotype might result by change

of gene expression level, whether directly by duplication or deletion of a gene or indirectly through position effects or downstream pathway and regulation networks (Dermitzakis and Stranger 2006; Reymond et al. 2007).

The first evidence that CNV could influence phenotype was via genomic re-arrangements (Inoue and Lupski 2002). The mechanism of the CNV's impact on the genome was considered as the result of a change in gene dosage changing (Kleinjan and van Heyningen 2005; Stankiewicz and Lupski 2006). There are two kinds of CNVs, losses (deletions or null genotypes) and gains (insertions or duplications). Deletions are known to be biased away from genes (Conrad et al. 2005), as a result of selection. But some kinds of diseases can be caused by deletion of a gene (Figure 1.2) or upstream of the gene (McCarroll et al. 2008) as in Figure 1.3. On the other hand, with an increase of copy number, the expression level shows four kinds of different trends, positive correlation, negative correlation, absence of correlation and multiple trend correlation (Guryev et al. 2008; Henrichsen et al. 2009b). This suggests there should be some kind of mechanisms for CNVs to influence the expression of the gene. There are six kinds of effects of CNV gains reported to be possible as shown in Figure 1.4 (Henrichsen et al. 2009a).

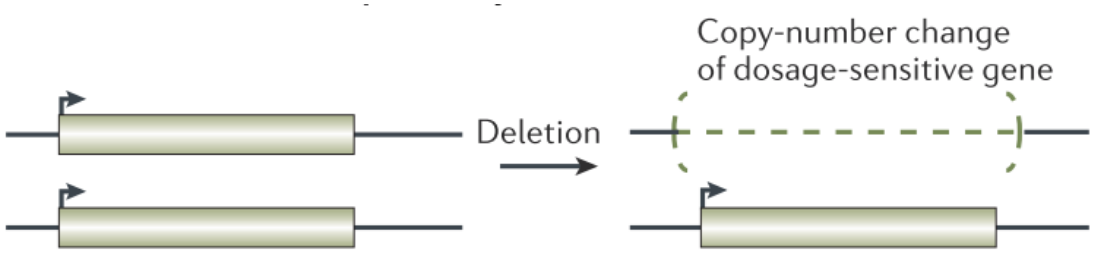


Figure 1.2 Deletion of gene causes change of gene dosage (Feuk et al. 2006).

The clear gray bar represents genes and the arrow represent direction of sequence (Reprinted, with permission, from Nature Reviews Genetics, Vol.7©2006 by Nature Publishing Group; permission license Number:4123841441616).

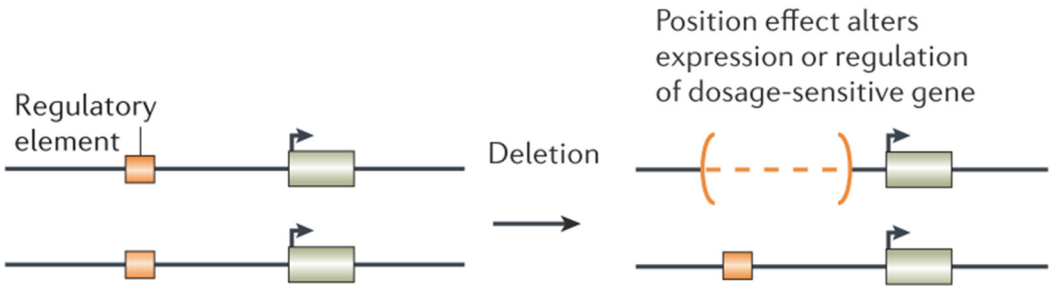


Figure 1.3 Deletion of upstream gene causes change of gene dosage (Feuk et al. 2006)**.**

The clear gray bar and orange box represents genes and regulatory. The arrow represent direction of sequence (Reprinted, with permission, from Nature Reviews Genetics, Vol.7©2006 by Nature Publishing Group; permission license Number:4123841441616).

1.3.3 Molecular mechanism of formation of CNV

So far, the formation of CNV is considered to be the result of errors occurring during DNA repair. One of the best recognized hypothesis is non-allelic homologous recombinations (NAHR) (Hastings et al. 2009). Homologous recombination is a kind of genetic recombination in which DNA sequences are exchanged between two similar or identical DNA sequences. This process happens during mitosis and is also a very important process for DNA repair. When these DNA sequences are not alleles, it is called NAHR. This mechanism leads to chromosomal structural changes because one DNA sequence become long (duplication) while the other become short (deletion) (Figure 1.5).

1.3.4 Methods for prediction of CNV

1.3.4.1 Array comparative genomic hybridization (aCGH)

Array comparative genomic hybridization (aCGH) was invented by Kallioniemi et al. (1992) to detect the differences between the chromosomal complements of solid tumor and normal tissue in 1992. After that, Solinas-Toldo et al. (1997). improved this technology to microarray rather than the traditional metaphase chromosome preparation.

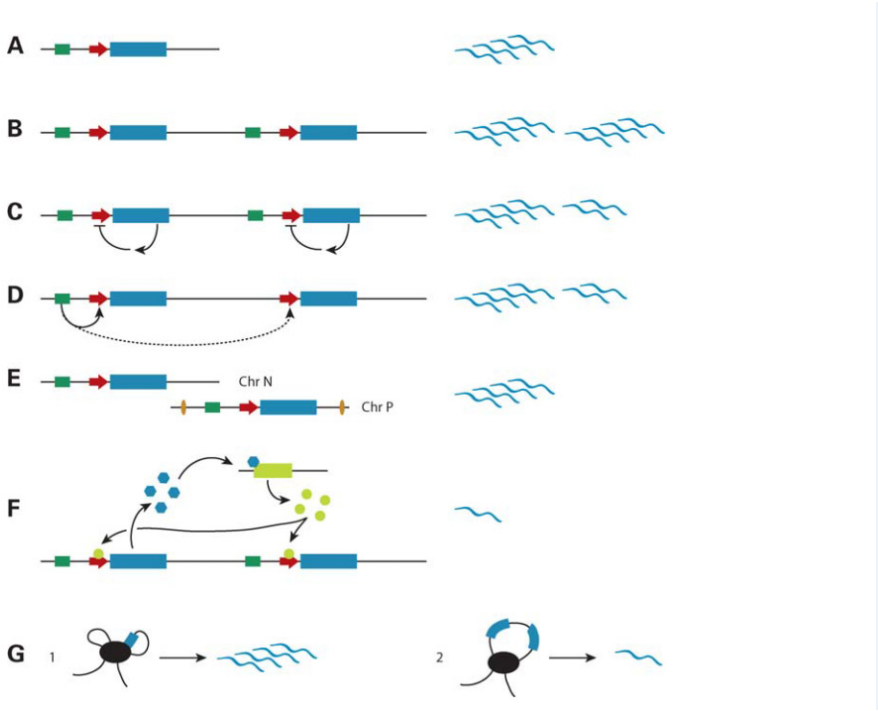


Figure 1.4 Duplication scenarios and their influence on expression (Henrichsen et al. 2009a)**.**

(A) Single copy gene locus (normal). The gene intron-exon region (blue box), the gene promoter (red arrow) and its enhancer (green box) are shown. Transcript levels are indicated schematically on the right. (B) Complete tandem duplication including the regulatory region. (C) Complete tandem duplication including the regulatory region of a gene under a compensatory mechanism. A negative feedback loop reduces the second copy of the gene, which is expressed at a lower level. (E) Complete non-tandem duplication including the regulatory region. The duplicated locus maps to another chromosome region where a different chromatin context, insulators (yellow ellipses), modifies its expression level. (F) Immediate early gene model. In the presence of a duplication, the concentration of the CNV gene product (blue hexagons) is sufficient to induce a repressor (light green box), the product of which (light green disks) blocks the expression of the CNV gene. (G) A tandem duplication (2) physically impairs the access of the CNV genes copies to the transcription factory where it should be transcribed (1).

A year later, Pinkel et al. (1998). used this technology to find DNA CNV firstly in 1998. The principle of the original CGH was to *in situ* hybridize differentially labelled (with fluorescent dye) test and reference genomic DNA to metaphase chromosome preparation and measure fluorescence ratios along the chromosome length for an idea about relative CNV. Current-day

CGH arrays employ bacterial artificial chromosome, cosmid or cDNA clones, or long synthetic oligos, instead of metaphase chromosomes, in order to probe either genome-wide or unique CNV. Until now, there were 27 papers published on the use of this technology for CNV detection in animals (Table 1.1).

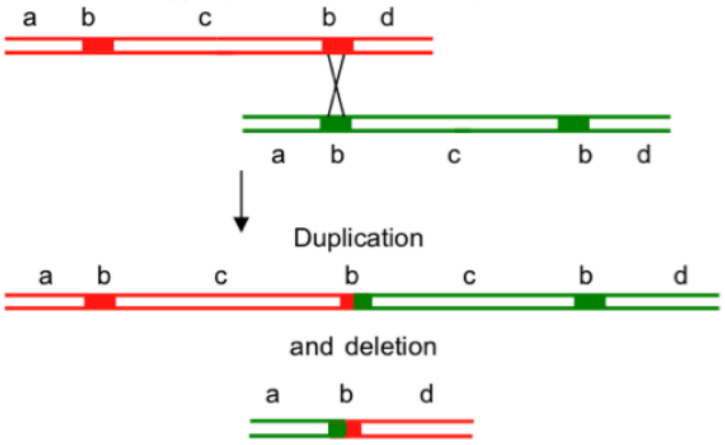


Figure 1.5 Non-allelic homologous recombination.

Red line and green line represents two chromosomes. The letters, a, b, c, d, represents homologous regions. A non-allelic homologous recombination happened on b region and lead to one chromosome becomes longer (duplication) and the other become short (deletion). This process results in CNV. (Reprinted, with permission, from Nature Reviews Genetics, Vol.10©2009 by Nature Publishing Group; permission license Number:4293870364435).

1.3.4.2 SNP microarray

Copy number varation (CNV) detection using SNP microarray is the most popular method in studies involving animals. There are about 3 million SNP distributed throughout the human genome that are ideal genetic markers to detect CNV. SNP microarray is a chip that has more than 500,000 SNP probes to detect SNPs at a time. By calculating the signal of a SNP, it is possible to predict CNV (see section 1.4.1). Until now, there are 51 published papers that are based on this technology for CNV studies in animals (Table 1.1).

1.3.4.3 Whole genome sequencing and PCR

By whole genome sequencing, the DNA sequence of an individual can be identified directly and compared with a reference genome database. Currently, there are three next generation sequencing (NGS) platforms. They are Roche GS FLX sequencer, Illumina Genome Analyzer and ABI SOLID sequencer. This method can produce the most precise CNV map (Kidd et al. 2008; Tuzun et al. 2005). Until now, only 9 published studies in animals were based on this technology (Table 1.1).

The CGH array was the first method to be used for CNV detection in animals. However, because of its relative low resolution (40 kb-several megabase) (Carter 2007), many miniature variation could be lost. With the development of array technology, SNP arrays have become more popular than CGH arrays based on their high resolution (minimum resolution for CNV detection is 10-40 kb) (Carter 2007). However, because the SNPs are not well-distributed across the whole genome, any variation which happens in an area without SNPs cannot be detected. Currently, all SNP microarrays nowadays are designed based on biallelic phase. Some kinds of variation which are not biallelic will be missed. Although NGS provides the

highest resolution among the three methods, it is very expensive. Therefore, it is not often the method of choice for large population research. SNP microarray and NGS were chosen for CNV detection and validation in this study.

1.3.5 Current research on CNV

1.3.5.1 Studies involving human disease and traits

Since CNVs were found, most studies about CNV have been done in human and associated with disease or phenotypes. As of February, 2018, the Database of Genomic Variants

(http://dgv.tcag.ca/dgv/app/home) lists 72 studies reporting over 6,359,956 CNVs on sample-level and 552,586 on a merged-level.

Many associations between CNVs and disease or phenotypes have been found. Firstly, inversion of CNVs has been reported to cause many kinds of diseases. Studies found one-third of parents of patients with Williams-Beuren syndrome have a 1.5-MB inversion at 7q11.23 (Osborne et al. 2001). Similarly, about half of the parents of patients with Angelman syndrome carry an inversion of 4 Mb at 15q12 (Gimelli et al. 2003), respectively. In Japan, most fathers of the patients with Sotos syndrome carry a 1.9-Mb inversion variant at 5q35 (Kurotaki et al. 2003).

CNVs are also associated with some kinds of autoimmune diseases. Low copy number of *FCGR3B* was associated with glomerulonephritis in the autoimmune disease systemic lupus erythematosus (Aitman et al. 2006). Low copy number of the *C4* gene was found as a risk factor for systemic lupus erythematosus (*SLE*) in European Americans, whereas high copy number of *C4* gene was protective (Yang et al. 2007). Strong associations between *FCGR3B* copy number and risk of systemic lupus erythematosus, microscopic polyangiitis and Wegener's granulomatosis were also found by Fanciulli et al. (2007).

Moreover, CNVs also effect susceptibility to infectious diseases. Today, the main research areas between CNVs and infectious disease in human are on malaria and HIV. The α-Globin is controlled by two genes, *HBA1* and *HBA2*, which are copy number variable. Different kinds of CNV in these two genes can lead to different phenotypes of erythrocyte as illustrated in Figure 1.6. Individuals with either 3 or 2 copies of α-Globin gene have the ability to protect against malaria because of changes in erythrocyte surface receptors (Hollox and Hoh 2014). Meanwhile, in HIV research, the copy number variation of the *CCL3L1* gene, which codes for a human immunodeficiency chemokine for HIV, has been found to be associated

with susceptibility to HIV. The more copies of chemokine, the lower the risk of HIV infection in humans (Gonzalez et al. 2005).

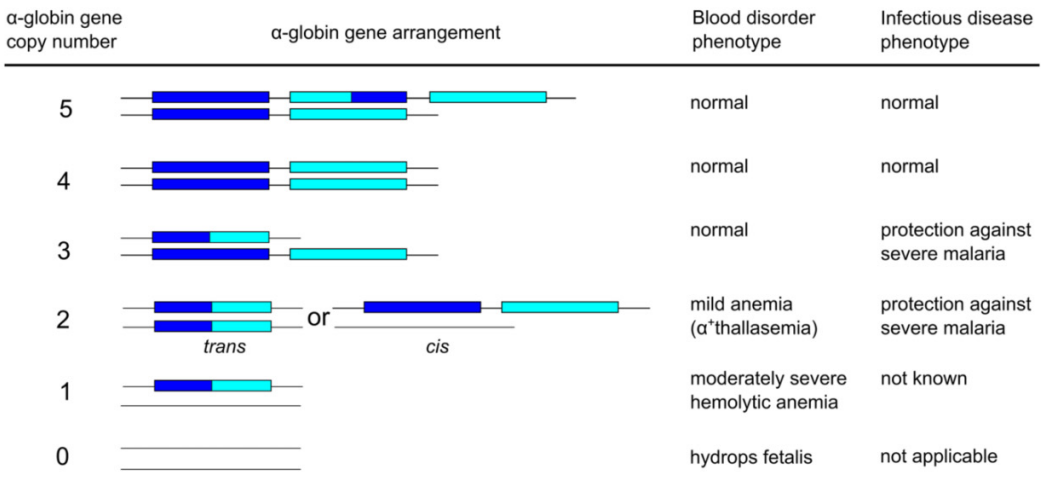


Figure 1.6 Copy number of α-Globin (HBA) and different clinical phenotypes.

Different observed diploid copy numbers of HBA are shown in descending order, together with the schematic gene arrangement (dark blue representing α-1-Globin and pale blue representing α-2-Globin), and the blood disorder and infectious disease phenotypes of each copy number (Hollox and Hoh 2014) (Reprinted, with permission, from Human genetics, Vol.133©2014 by Springer; permission license Number: 4123890666723)

1.3.5.2 Studies in domestic animals

There is an increasing interest in research about CNV effects in domestic animals - cattle, sheep, goat, pig, horse, dog, chicken, turkey and duck. The study methods include aCGH, SNP array and whole-genome sequencing (Table 1.1). Since CNV was identified by Liu et al. (2008) in cattle in 2008, firstly with 25 CNVs, more than 16 studies have been done in cattle. Three main methods, aCGH, SNP array and whole genome sequencing, have all been adopted to identify CNVs in cattle. Matukumalli et al. (2009) firstly used a SNP chip to find CNV in a wide array of breeds. Then two detailed bovine CNV maps were made by Bae et al. (2010) and Fadista et al. (2010) using Bovine SNP50 BeadChip and custom aCGH array, respectively. One year later, Stothard et al. (2011) identified CNV in cattle via whole-genome sequencing for the first time. After that, with the improvement of SNP microarrays, several CNV studies have beenundertaken (Table 1.1).

Table 1.1 Studies on CNV detection in domestic animals.

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Cattle	556	42	960.6	394.8	22.9-11,050.6	BovineSNP50 BeadChip (cnvPartition)	(Matukumalli et al. 2009)
	265	368	171.5	128.3	50-200	BovineSNP50 BeadChip (cnvPartition)	(Bae et al. 2010)
	20	304	72.3	16.7	1.7-2,000	Bovine 2.1M aCGH arrays	(Fadista et al. 2010)
	90	177	159	89	18-1,260	Bovine 385k aCGH arrays	(Liu et al. 2010a)
	539	682	204.9	131.1	32.5-5,569	BovineSNP50 BeadChip (PennCNV)	(Hou et al. 2011)
	912	418				BovineSNP50 BeadChip (PennCNV)	(Seroussi et al. 2010)
	2,047	99	234.8	151.7	27.01-1,310	BovineSNP50 BeadChip (PennCNV, cnvPartition and GADA)	(Jiang et al. 2012)
	6	1,265	49.1	23.63	10.02-510.9	Next generation sequencing	(Bickhart et al. 2012)
	9	51	213-335			Bovine 385k aCGH arrays	(Kijas et al. 2011)
	2	790	4.163	3.171	1.841-28.029	Whole-genome sequencing	(Stothard et al. 2011)
	1	46		25.812	3.170-595.739	CNVseq+Nimblegen6 .3M+Illumina SNP770 k	(Zhan et al. 2011)

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Cattle	674	3,346	42.65	15.794	1.018-5,500	Btau_4.0 BovineHD SNP array	(Hou et al. 2012a)
		3,438	47.37			UMD3.1 BovineHD SNP array	
	96	367	96.23	50.69	10.76-2,806.42	Bovine high-density(770K) SNP arrays PennCNV	(Jiang et al. 2013)
	2,654	402	1,240	782	53-10,552	BovineSNP50 BeadChip QuantiSNP PennCNV	(Cicconardi et al. 2013)
	29	605				Nimblegen3x720K aCGH array	(Zhang et al. 2014b)
	32	6,811 (CNVs)	2.732			Next generation sequencing	(Shin et al. 2014)
	6	425	35.07	18.56		BovineHD Genotyping SNP BeadChip	(Quanwei Zhang et al. 2015)
	492	334			30-1,000	Illumina BovineSNP50 Beadchip	(Wang et al. 2015)
	792	263	134.7	61.95	10.18-1760	Illumina Bovine HD SNP BeadChip (770k)	(Wu et al. 2015)
	1,160	710				BovineSNP50 assay	(Gurgul et al. 2015)
	1,481	861	50.7		1.1-1,400	Illumina Bovine High-Density (HD) SNP BeadChip	(Sasaki et al. 2016)
	300	257	48.4			BovineHD SNP array	(Xu et al. 2016)
	75	1,853				Sequencing	(Bickhart et al. 2016)

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Cattle	1,725				20.1-3,810	Illumina Bovine HD Genotyping SNP Bead Chip Illumina HiSeq2000	(da Silva et al. 2016)
	242	252				Illumina BovineHD SNP BeadChip	(Durán Aguilar et al. 2017)
Sheep	11	135	77.6	55.9	24.6-505	Bovine 385k aCGH	(Fontanesi et al. 2011)
	329	238	253.57	186.92	13.66-1,300	Ovine SNP50 BeadChip PennCNV	(Liu et al. 2013)
	160	111	123.84	100.53		Ovine SNP50 BeadChip	(Youji Ma et al. 2015b)
	5	51	304.86		52-2,000	1.4 M aCGH	(Hou et al. 2015)
	36	3,488	19		1-3,600	2.1M aCGH	(Jenkins et al. 2016)
	120	490	165.39	133.17	100.11-804.18	Illumina Ovine SNP 600 BeadChip	(Zhu et al. 2016)
	385	749	189	118	15.3-6,600	OvineSNP50K	(Yan et al. 2017a)
		464	305.5	218.1	11.4-2,108.8		
		104	1521.3	395.4	87-12,093.7		
	48	1,296	92.7		1.2-2,300	Illumina OvineSNP 600 K BeadChip	(Ma et al. 2017)
	2,254	24,588				OvineSNP50K	(Yang et al. 2018)

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Pig	12	37	9.32	6.89	1.7-61.9	Porcine 385k aCGH	(Fadista et al. 2008)
	55	49	754.6	170	44.65-10,715.82	Porcine SNP60 BeadChip (PennCNV, cnvPartition and GADA)	(Ramayo-Caldas et al. 2010)
	474	382	250.7	142.9	5.03-2702.7	Porcine SNP60 BeadChip (PennCNV)	(Wang et al. 2012a)
	12	259	65.07	98.74	2.3-1,550	720 K array CGH (aCGH)	(Li et al. 2012)
	1,693	565	247.55	252.71	50.39-8,102.06	Porcine SNP60 BeadChip (PennCNV)	(Chen et al. 2012)
	14	63	158.37	97.85	3.2-827.21	Infinium II Multisample SNP assay (PennCNV)	(Wang et al. 2013a)
	117	1,928 (CNVs)	5.23	3	0.12-175.5	Illumina HiSeq technology	(Rubin et al. 2012)
	585	660	1,880			Porcine SNP60 BeadChip (GADA)	(Wang et al. 2013b)
		505	210			Porcine SNP60 BeadChip (PennCNV)	
		966	1,050			Porcine SNP60 BeadChip (QuantiSNP)	
		60	2,570			Porcine SNP60 BeadChip (cnvPartition)	
		249		845.98	29.20-27,290	Total	

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Pig	16	3118	12.74	10	6-96	Illumina HiSeq platform	(Paudel et al. 2013)
	288	216				Porcine SNP60 BeadChip (QuantiSNP)	(Fowler et al. 2013)
		27				Porcine SNP60 BeadChip (cnvPartition)	
	12	1,344	35.56	11.11	3.37-1,319	Custom-designed 2.1 M array (aCGH)	(Wang et al. 2014a)
	13	3,131	32.8			Illumina HiSeq 2000	(Jiang et al. 2014a)
	150	5			192.97-4,892.37	Porcine SNP60 BeadChip (cnvPartition)	(Long et al. 2014)
	302	348	443.24	170.77	4.93-12,410	Porcine SNP60 BeadChip (PennCNV)	(Wang et al. 2014c)
	297	170		180.3	25.2-1,700	Porcine SNP60 BeadChip (PennCNV)	(Schiavo et al. 2014)
	223	65	148.99		3.06-1,007	Porcine SNP60 BeadChip (PennCNV)	(Fernandez et al. 2014)
	96	105				Porcine SNP60 BeadChip (PennCNV)	(Dong et al. 2015)
	611	165			5.03-652.41	Illumina PorcineSNP60 BeadChip V2	(Zhou et al. 2016)
	252	44,511 (CNVs)				Illumina HiSeq2000	(Wang et al. 2017)

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Horse	1	282 CNVs			3.74-4,840	Whole-genome sequencing	(Doan et al. 2012a)
	16	775	99.4	5.3	197-3,500	Custom-designed aCGH	(Doan et al. 2012b)
	96	239	79.67			Porcine SNP60 BeadChip (PennCNV)	(Dong et al. 2015)
	6	367	35.07	18.56		Bovine SNP HD(PennCNV)	(Quanwei Zhang et al. 2015)
	854	50 (CNVs)	388.892	293.244	0.516-978.535	Equine SNP50 beadchip (PennCNV, cnvPartition and QuantiSNP)	(Metzger et al. 2013)
	447	478				Equine SNP50 beadchip (PennCNV)	(Dupuis et al. 2013)
	6	353	38.49	13.1	6.1-1,450	aCGH chip designed by Roche NimbleGen	(Wang et al. 2014b)
	4	1246				Re-sequencing data	(Park et al. 2014)
	96	122	1,552		199-2,344	Equine 70K SNP genotyping array	(Kader et al. 2016)
Dog	17+1 breeds	3,583 CNVs				aCGH chip designed by NimbleGen	(Nicholas et al. 2009)
	9	60				aCGH chip designed by NimbleGen	(Chen et al. 2009b)
	61	403				Custom aCGH chip	(Nicholas et al. 2011)

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Dog	50	430 (CNVs)				Custom-designed 2.1 M array (aCGH)	(Berglund et al. 2012)
	7	5 (CNVs)				NimbleGen custom 720K aCGH canine whole-genome array	(Jung et al. 2013)
	359	72		194.559		Illumina 170 K CanineHD SNP array (PennCNV, QuantiSNP)	(Molin et al. 2014)
	23	1,161 (CNVs)				Custom 720K probe aCGH chip	(Ramirez et al. 2014)
Chicken	2	12 (CNVs)	127	90	30-300	Chicken 385k aCGH arrays	(Griffin et al. 2008)
	10	96 (CNVs)	168.1	43.6	10.34-14,102.44	Chicken 385k aCGH arrays	(Wang et al. 2010)
	18	130	25.7	14.43	6.20–649.12	Agilent 400k array CGH	(Wang et al. 2012b)
	746	209				Chicken 60K SNP array (PennCNV)	(Jia et al. 2013)
	22	308	35.1	14.6	5.8-2,000	Agilent 400k aCGH	(Tian et al. 2013)
	64	1,556				Agilent 244K chicken aCGH	(Crooijmans et al. 2013)
	12	273 (CNVs)				Agilent 244K chicken aCGH	(Abernathy et al. 2014)
	203 lean	271	148.77	107.81	6.23-932.14	Chicken 60K SNP array (PennCNV, cnvPartition)	(Zhang et al. 2014a)

Species	Number of animals	CNVR				Platform	References
		Number of CNVRs or CNVs	Mean size (kb)	Median size (kb)	Size range (kb)		
Chicken	272 fat	188	163.43	99.81	0.33-1,442.99	Chicken 60K SNP array (PennCNV, cnvPartition)	(Zhang et al. 2014a)
	12	8,840				Next-generation sequencing	(Yi et al. 2014)
	1,310	137		199.4	11-3,034	SNP 60 K PennCNV	(Zhou et al. 2014)
	10	281	45.6	25.0		385 K (aCGH)	(Han et al. 2014)
	6	3,241	8.4		1-543.5	Illumina HiSeq 2000	(Yan 2015)
	554	383				Chicken 60K SNP BeadChip	(Rao et al. 2016)
	94	564				Axiom®Genome-Wide Chicken Genotyping SNP Array (Affymetrix)	(Strillacci et al. 2016)
	256	1,216				600K SNP chip array	(Gorla et al. 2017)
Duck	2	32	281	50	2.8-515.181	Chicken 385k aCGH arrays	(Skinner et al. 2009)
Turkey	1	16	179	90	30-900	Chicken 385k aCGH arrays	(Griffin et al. 2008)
Goat	10	127	90.3	49.5	24.6-1,070	Bovine 385k aCGH	(Fontanesi et al. 2010)

Research to date shows that different platforms and CNV discovery algorithms can produce very different results. The results from different studies overlap partially (Cicconardi et al. 2013; Hou et al. 2012a). Zhang et al. (2014b) found that there were a few CNV overlapping among three different relative groups (taurine, yak and buffalo).

Until now, there are 15 studies reported about CNV inpig. Fadista et al. (2008) published the first CNV identification in pig by using aCGH and found 37 CNVRs. Ramayo-Caldas et al. (2010) used the Porcine SNP60BeadChip and found 49 CNVRs. The first whole-genome sequencing in pig for CNV was published by Rubin et al. (2012) who found 1928 CNVs. Wang et al. (2013b) used four different kinds of algorithms to deal with chip data and got four different kinds of results about CNV calling. The result showed that the CNVRs identified by QuantiSNP were maximum (966) while those identified by cnvPartition were the least (60). There are a few studies about sheep and goat. There are eight papers on sheep (Table 1.1) and one on goat (Fontanesi et al. 2010). Those studies used either aCGH or SNP arrays, none were based on whole-genome sequencing.

Other mammals, such as dog and horse, have also been studied (Table 1.1). Chen et al. (2009b) finished the first map of CNV in dogs by using aCGH and found 155 CNVs in 60 CNVRs while Doan et al. (2012b) finished the first map of CNV in horse. After that, Jung et al. (2013) identified 5 CNVs in 7 cloned dogs via the NimbleGen custom 720K canine whole-genome array platform (aCGH). This was also the first time CNVs were found in cloned animals and gave a new way to understand formation mechanisms of genetic variants.

Chicken, duck and turkey have also been studied for CNVs. The studies done by Griffin et al. (2008) and Skinner et al. (2009) support the hypothesis that avian genomes contain fewer CNVs than mammalian genomes and that genomes of evolutionarily distant species share regions of CNV. Wang et al. (2010) finished the first map of CNV in chicken in 2010 and found 96 CNVs.

Through CNV studies, some interesting traits have been associated with CNV, which are mainly focused on pigment and coat colour morphology. The coat colour of several animals, (horse, pig and sheep) has association with CNV. Giuffra et al. (1999, 2002) found that the

duplication of a 450-kb fragment encompassing the KIT gene can lead to white coat colour of pigs. Rosengren et al. (2008) found a 4.6 kb duplication in intron 6 of the STX17 gene, which affected the hair depigmentation of horse. A 100-kb duplication of the ASIP gene associated with white coat in sheep (Fontanesi et al. 2009). The morphology of animals also can be effected by CNV. Elferink et al. (2008) discovered a 176 kb duplication containing the PRLR and SPEF2 genes affected feather growth in chickens. Wright et al. (2009) found a massive duplication event that altered SOX5 expression, led to pea-comb phenotype in chickens. Research about production traits has also been done. Hou et al. (2012b) attempted to find potential mechanisms contributing to differences in residual feed intake by analysing CNV in Holstein cows. Hou et al (2012c) studied the association between CNV and resistance or susceptibility to nematodes in Angus cattle (Hou et al. 2012c).

1.3.6 Summary

Several studies have been undertaken during the past decade in humans as well as animals and CNV has shown its potential as a new genetic marker for primarily disease and morphological traits. Different kinds of platforms, such as aCGH, SNP microarrays and NGS, have been used for detection of CNV in those studies. These platforms have specific advantages and disadvantages. SNP microarrays have been most popular for CNV detection due to low cost and relatively higher resolution than aCGH. Although NGS offers high resolution, they are not that popular in animal studies due to high cost. Compared to other domestic animal species, CNV studies were rare in sheep (eight studies) and goat (only one study).

1.4 Details of CNV detection platforms

1.4.1 SNP microarray

1.4.1.1 Background of SNP arrays

There are two main commercial SNP chip manufacturers – Illumina (SanDiego, CA, USA) and Affymetrix (Santa Clara, CA, USA). Affymetrix deals mostly with chips for human and Illumina manufactures SNP chips for different animal species. Details of SNP detection using the case of Illumina chip and different algorithms for detection of CNV from SNP genotypes are reviewed below.

1.4.1.2 Operational principle of the Illumina microarray

The Illumina Bead Chip works on a randomly ordered BeadArray technology, which was invented by Professor David Walt and colleagues at Tufts University (Michael et al. 1998; Walt 2000). This technology takes advantage of the intrinsic structure of the optical fibers (Oliphant et al. 2002). Each fiber has a light-conducting inner core that is surrounded by a cladding of different refractive index. The core can be chemically etched at a different rate from its surrounding cladding. By treating the polished end of an optical fiber bundle with acid, an array of micro wells is generated. The geometry and dimensions of the array are determined by the physical specifications of the optical fiber and are chosen so that one bead can fit in each well in the array. Once a labeled target nucleic acid is hybridized to beads in the array, a fluorescent signal can be generated by making use of the optical properties of the fiber. An excitation beam is guided to the bead through the fiber bundle, and emitted fluorescence is guided back up the fiber, allowing the array to be imaged at the opposite end of the optical fiber bundle. There are 1536 bead types in Illumina BeadArray, each one with a unique capture sequence to combine with target DNA (Oliphant et al. 2002). Each chip can test 16 samples at a time and include about 50,000 beads (bead types and number depend on the type of microarray and could change). The workflow of Illumina bead array is depicted in

Figure 1.7. Each SNP probe is a 50-mer sequence complementary to the sequence adjacent to the SNP site. A single base extension of the probe binds to the complementary base at the SNP site and results in appropriate color (red in case of A or T, and green in case of C or G) (Figure 1.8). In a DNA sequence, one point (base) can possibly mutate to the other 3 kinds of bases, for example A to either T, G or C. However, based on the biallelic natures of SNP, normally the SNP just mutates to only one kinds (C➔T/G➔A) (Wang et al. 1998). Thus, three types of genotypes are possible in a population (Figure 1.9).

1.4.1.3 Development of SNP microarrays for sheep

So far, there have been three kinds of SNP microarrays used in sheep. They are the OvineSNP50 Genotyping BeadChip, the Low Density (5K) Ovine SNP Chip, and OvineHD

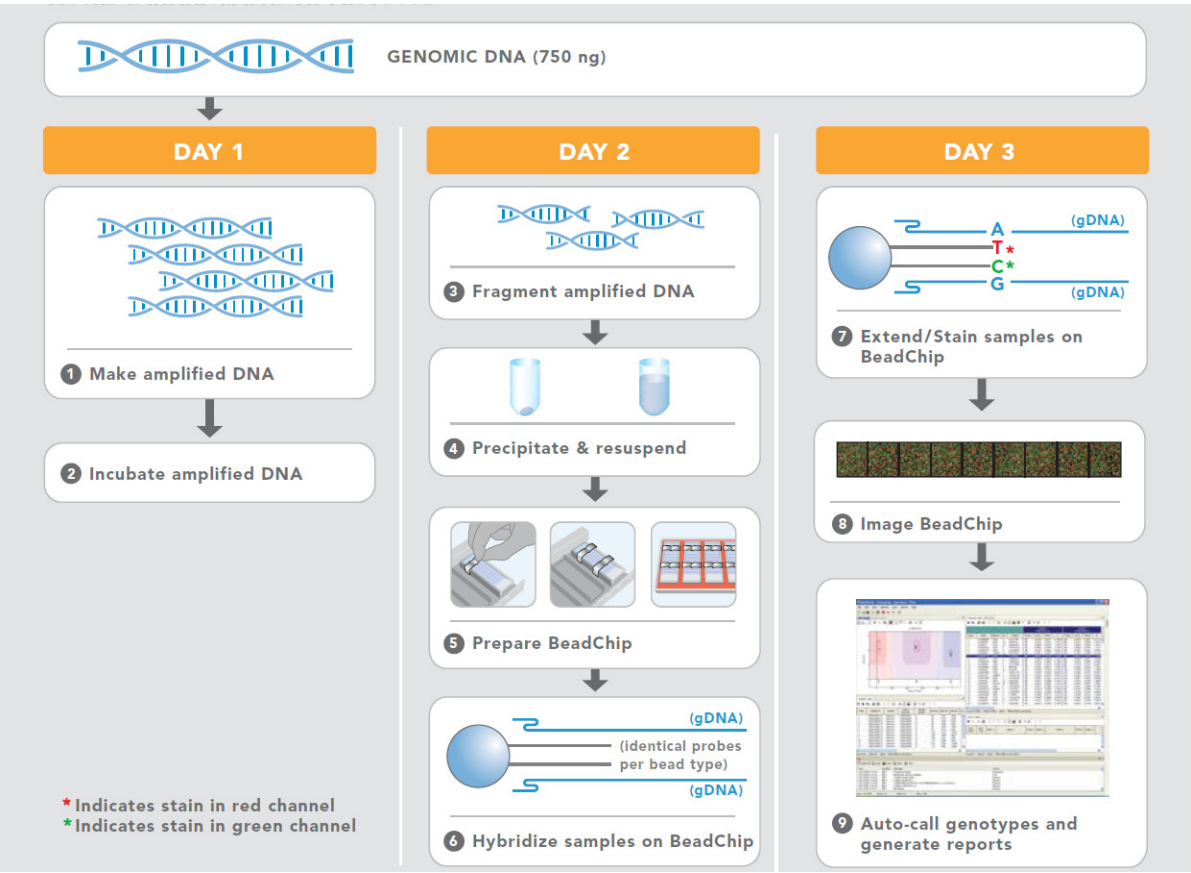


Figure 1.7 Illumina bead chip workflow (Illumina 2012).

The first SNP microarray in sheep was the OvineSNP50 Genotyping BeadChip, which was designed in 2007 by the International Sheep Genomics Consortium (ISGC) using the Roche 454 FLX sequencing (0.5X) based on 6 female sheep (Oddy et al. 2007). In total, 595,000 SNPs were identified and after quality control the chip desighed 49,034 useful SNPs. The second SNP microarray was the Low Density Ovine SNP Chip (5K), which used a subset of probes from the OvineSNP50 Genotyping BeadChip. This chip only carries 5, 998 SNPs (Anderson et al. 2012), consequently cost less than the OvineSNP50 Genotyping BeadChip.

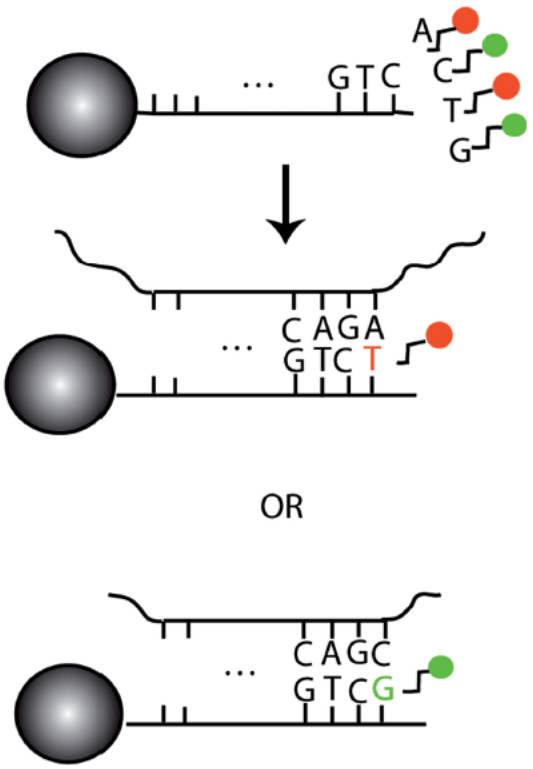


Figure 1.8 The principle of base detection.

The principle of base detection. Four kinds of nucleotide labeled by 2 kinds of color. Attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site. The single-base extension (T or G) that is complementary to the allele carried by the DNA (A or C, respectively) then binds and results in the appropriately-colored signal (red or green, respectively). For both platforms, the computational algorithms convert the raw signals into inferences regarding the presence or absence of each of the two alleles (LaFramboise 2009) (Reprinted, with permission, from Nucleic acids research, Vol.37©2009 by OXFORD UNIVERSITY PRESS; permission license Number: 4287480202675).

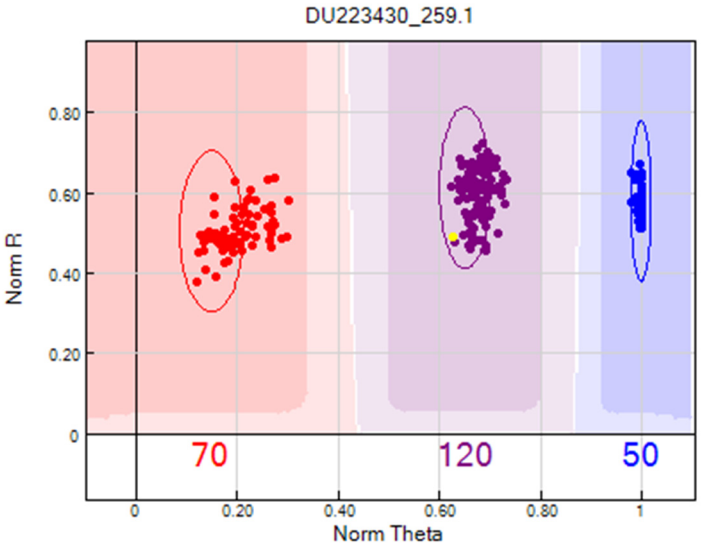


Figure 1.9 The genotyping results of one SNP point in a group of samples from the current study (Chapter 2).

Each point represents an animal. X-axis and Y-axis represent Norm theta and Norm R, respectively. The genotypes (AA, AB, BB) are represented by red, purple and blue colors.

The OvineHD Array (600K) was designed in 2012 by the International Sheep Genomics Consortium (ISGC) based on Whole Genome Sequencing Project (Kijas et al. 2012). This project used Illumina paired end sequencing based on 75 individuals and aimed to produce 50 million SNPs which is the foundation for the design of the OvineHD Array. Currently, this chip has more than 600,000 SNPs.

1.4.1.4 CNV detection workflow in SNP Microarray

Typically, there are 5 steps in CNV workflow - process raw data and generate log ratio, quality assurance, detection of copy number variants, statistical analysis of CNVs and make sense of findings (Figure 1.10). In the first step, raw data measured as light intensity is normalized and processed to generate log ratios, so as to make comparisons between animals possible. However, any study using a SNP microarray is accompanied with significant bias such as a batch effect, experimental variability and the signal-to-noise properties of each sample. So the data has to be filtered by quality assurance (control). After that, the light

intensity signal can be used for CNV detection using an appropriate algorithm. Then, by statistical analysis, possible associations between CNV and phenotypes can be tested. The final step is to use bioinformatics on the CNVs found to be significant, so as to identify the biochemical and molecular significance of the genes located in the region.

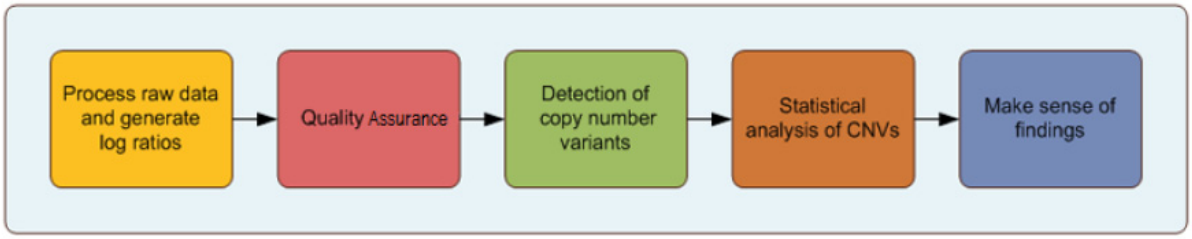


Figure 1.10 CNV work flow.

(CNV Univariate analysis Tutorial. SNP & Variation Suite Manual v8.7. Copyright © 2017 Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com.)

1.4.1.5 Algorithms for CNV detection based on SNP microarray

Based on the working mechanism of SNP microarray, two alleles of one SNP locus are designed as A or B whose normalized intensities are *a* and *b*, respectively. In CNV detection, a variable, *R*, is used to measure the combined signal intensity of two alleles ($R = a + b$). Log R ratio (LRR) is defined as $\log2(R_{observed}/R_{expected})$, in which $R_{expected}$ is measured from reference samples (Li and Olivier, 2013). Besides, another variable called B allele frequency (BAF) is used in SNP micorarry, which is the normalized measure of relative signal intensity ratio of two alleles. The meaning of BAF could be different based on SNP chip manufacturers. The microarray used in this thesis was made by Illumina. Therefore, the meaning of BAF is the proportion of hybridized sample that carried the B allele.) Most methods that use SNP arrays to detect CNVs use both LRR and BAF. More than 12 algorithms have been published (listed in Table 1.2) for CNV detection using Illumina microarrays. Most of them use hidden Markov models (HMMs) and segmentation algorithms.

Table 1.2 Summary of popular algorithms for CNV detection using Illumina SNP microarrays

Algorithm	Model	Reference
PennCNV	Hidden Markov Model (HMM)	(Wang et al. 2007)
cnvPartition	HMM	(Illumina 2017)
Golden Helix SNP & Variation Suite (SVS)	Optimal Segmentation	(Helix 2017)
QuantiSNP	Objective Bayes-HMM	(Colella et al. 2007)
GenoCNV	HMM	(Sun et al. 2009)
MixHMM	HMM	(Liu et al. 2010b)
CNstream	Heuristics and parametrical statistics	(Alonso et al. 2010)
cn.FARMS	Probabilistic latent variable model	(Clevert et al. 2011)
GADA	Compact liner algebra	(Pique-Regi et al. 2008)
CNVworkshop	Circular Binary Segmentation algorithm	(Gai et al. 2010)
R-Gada	Bayesian learning	(Pique-Regi et al. 2010)
NEXUS	Rank Segmentation, SNPRank Segmentation, FASST Segmentation and SNP-FASST Segmentation	(Darvishi 2010)
SCIMMkit	Two rounds of mixture likelihood-based clustering	(Zerr et al. 2010)

1.4.2 NGS

1.4.2.1 Operational principle of the Illumina NGS

The Illumina NGS is done based on Illumina flow cell (Figure 1.11) technology. Processing is done in four steps: library preparation, cluster generation, sequencing, and data analysis. Firstly, genomic DNA extracted from samples is broken down into smaller pieces by ultrasound, then adapters are ligated to ends of each fragment (Figure 1.12 left). Primers adhere to the adapters that have been planted in the flow cell channels (Figure 1.12 right). Next, a bridge PCR amplifies the DNA sequence (Figure 1.13). By using an alkali solution, the amplified DNA double strand is untied into two single strands as the template for further processing. A special sodium azide modified dNTP is used in the PCR so that in each cycle only one base pair can be added (Figure 1.14 right). Then, the whole flow cell will be

scanned to recode the fluorescent signal from each base pair just added (Figure 1.14 left). This process needs to be repeated several times. However, with the increasing number of base pairs added, the accuracy will decrease dramatically so that effective read length is about 150 bp, using the Illumina pair-end technology, the effective read length can be extended to 300 bp.

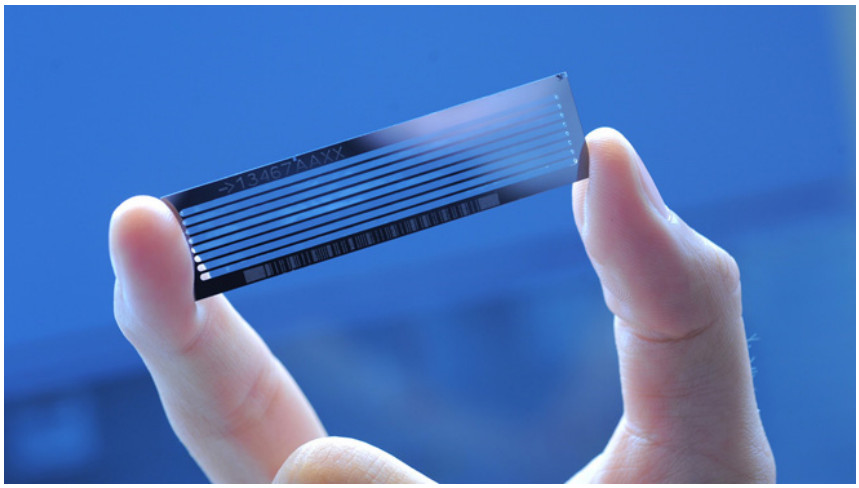
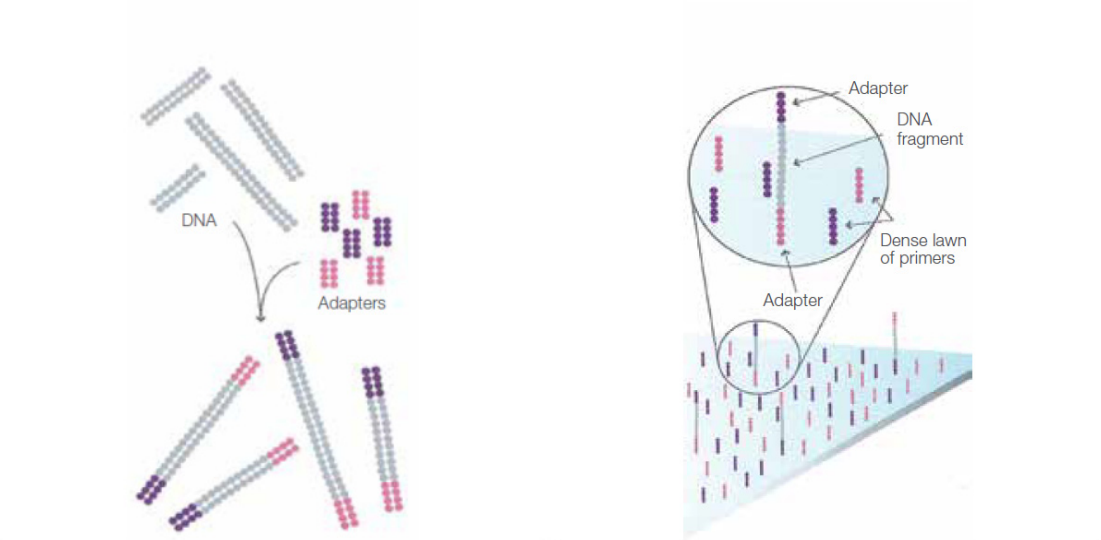


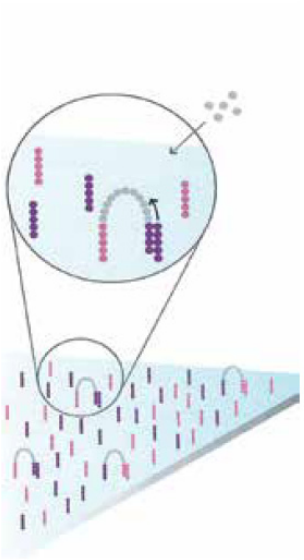
Figure 1.11 Illumina flow cell (Frank-Vinken-Institute 2013).



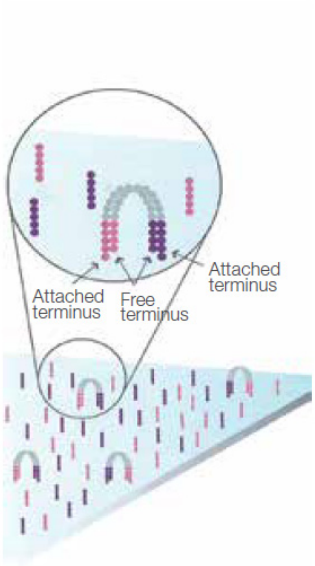
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Figure 1.12 Illumina NGS overview 1 (Illumina 2010).

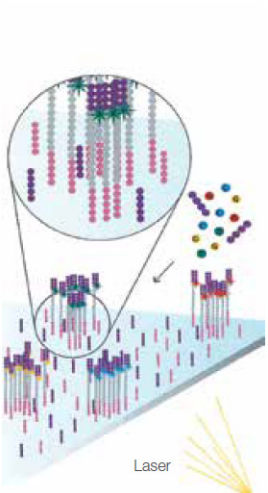


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Figure 1.13 Illumina NGS overview 2 (Illumina 2010).



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Figure 1.14 Illumina NGS overview 3 (Illumina 2010)

1.4.2.2 CNV detection workflow in NGS

Typically, there are 5 steps in CNV workflow based on NGS: obtain raw data, quality assurance and mapping, CNV calling, data analysis and making sense of findings. Firstly, raw data (the fluorescence signal) is normalized and processed to identify the type of base pair (A, T, C, G), this generates raw reads data. By mapping these reads to a reference genome assembly, a mapped file will be created for further study. The mapped sequence data is then used for CNV detection using an appropriate algorithm. Statistical analysis is used to find possible associations between CNV and phenotypes. The final step is to conduct bioinformatics studies on the CNV found to be significant so as to identify the biochemical and molecular significance of the genes located in the region.

1.4.2.3 Algorithms for CNV detection based on NGS

The algorithms for CNV detection based on NGS is totally different to those based on SNP microarray. So far, there are five strategies for CNV detection using NGS data (illustrated in Figure 1.15) (Zhao et al. 2013), which are paired-end mapping (PEM), split read (SR), read depth (RD), de novo assembly of a genome (AS) and combination of the above approaches (CB). More than 30 kinds of algorithms, using these five strategies, have been in use (Table 1.3).

1.4.3 Summary

In this thesis, two kinds of platforms, SNP microarray and NGS, were used for CNV detection and hence details regarding these two platforms have been reviewed. SNP microarray in sheep is available in three formats - Low Density (5k), 50K and High Density (600K). With the increasing number of probes, the resolution and cost goes up. More than 12 different algorithms have been developed for CNV detection based on SNP microarray data.

Meanwhile, 5 different strategies have been in use for CNV detection based on NGS data and more than 30 kinds of algorithms have been invented.

GWAS (genome-wide association study) is a statistic method based on the chi-squared test for case-control traits and F-test for quantitative traits to discover relationships between genotypes and phenotypes. The first successful GWAS was published in 2005 by Robert et al (Klein et al. 2005), two SNPs found to be significant with age-related macular degeneration. After that, thousands of individuals have been tested and more than 1,200 GWA studies have been done in 200 diseases and traits, and about 4,000 SNP-phenotype relationships found (Johnson and O'Donnell 2009) in humans.

GWAS can be done based on several kinds of genotype marker, SNP data being the most popular one. SNP have revolutionized the search for genetic markers for functional and disease traits in humans as well as animals. Several associations of SNP with different traits have been detected. However, in majority of the cases, they do not explain or account for the complete genetic variation seen in the traits.

CNV have recently been identified as contributing to genetic variation in production and disease traits. During the last decade, several studies in humans and animals have identified CNV polymorphisms and their association with quantitative as well as complex disease traits. Noteworthy, of the significant CNV relationships identified in livestock include: duplication of the agouti signaling protein gene, that results in white coat color in sheep (Norris and Whan 2008) and goat (Fontanesi et al. 2009) and duplication of CIITA, a trans-activator of MHC II, associated with nematode resistance in Angus cattle (Liu et al. 2010a). Hence, CNV could also be a useful genetic marker for GWAS.

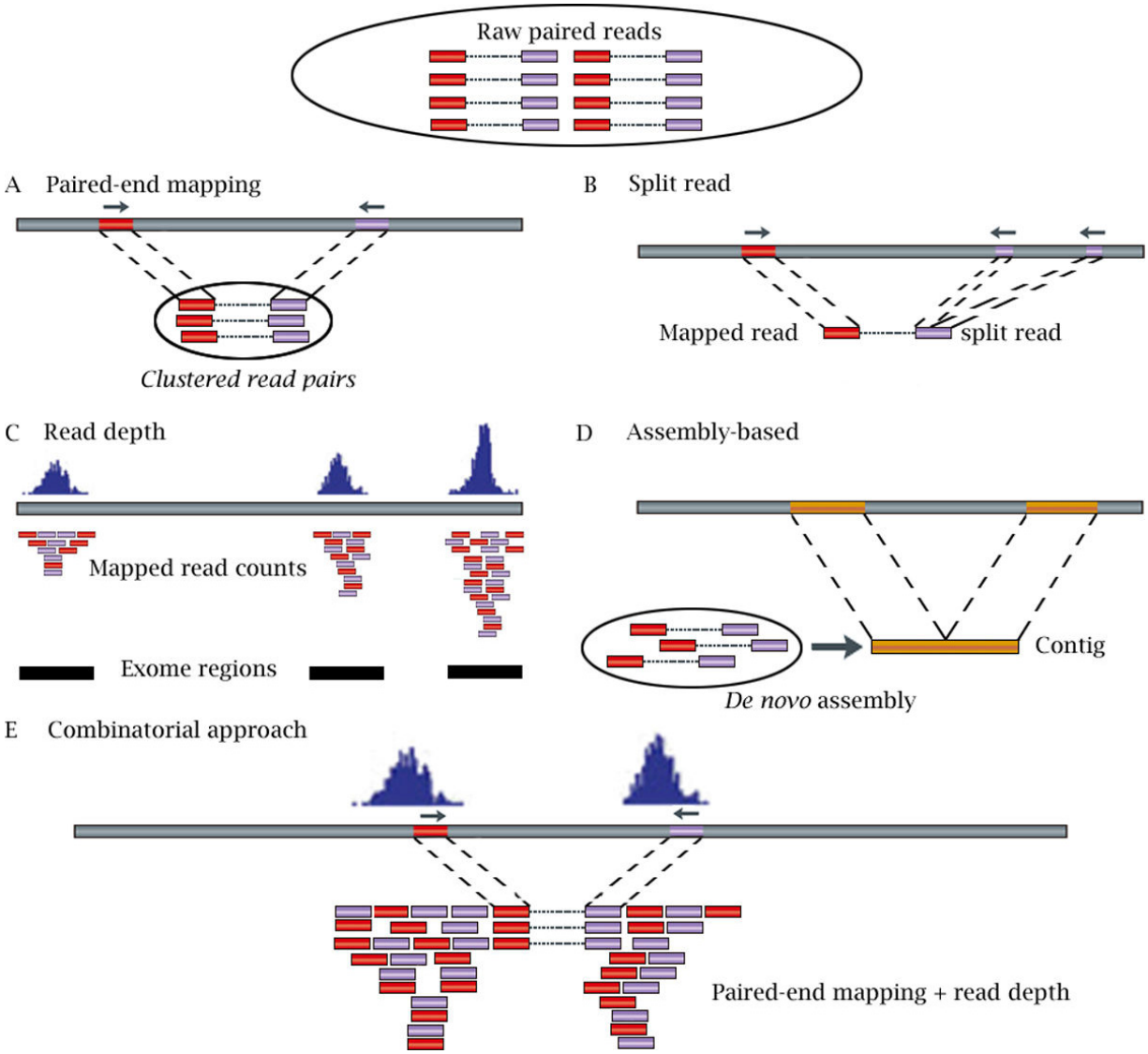


Figure 1.15 Five approaches to detect CNVs from NGS short reads.

A. Paired-end mapping (PEM) strategy detects CNVs through discordantly mapped reads. A discordant mapping is produced if the distance between two ends of a read pair is significantly different from the average insert size. B. Split read (SR)-based methods use incompletely mapped read from each read pair to identify small CNVs. C. Read depth (RD)-based approach detects CNV by counting the number of reads mapped to each genomic region. In the figure, reads are mapped to three exome regions. D. Assembly (AS)-based approach detects CNVs by mapping contigs to the reference genome. E. Combinatorial approach combines RD and PEM information to detect CNVs. (Reprinted, with permission, from BMC bioinformatics, Vol.14©2013 by SpringerOpen; permission agreement: All SpringerOpen articles are made available and publicly accessible via the Internet without any restrictions or payment by the user, reproduction of figures or tables from any article is permitted free of charge and without formal written permission from the publisher or the copyright holder https://www.springeropen.com/about/reprints-and-perm).

1.5 Selection signatures

Selective sweep is a kind of genetic phenomenon such that the genetic diversity around a mutation location will be reduced or eliminated (Figure 1.16). This occurs because the alleles around the mutation location are inherited to next generation due to linkage with positive selection occurring at the mutation location.

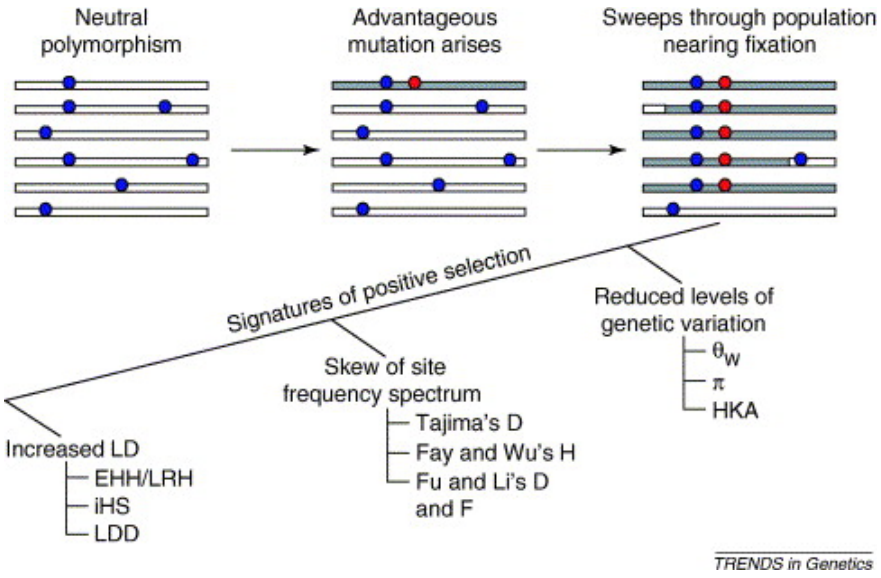


Figure 1.16 The formation of Selective sweep (Biswas and Akey 2006).

Blue point represents neutral mutation and red point represents advantageous mutation. The bars represent chromosomes. (Reprinted, with permission, from Trends in Genetics, Vol.22©2006 by Elsevier; permission license Number: 4123940988048)

Table 1.3 Summary of current algorithms for CNV detection based on NGS.

Algorithms	Strategies	Reference
BreakDancer	PEM-based	(Chen et al. 2009a)
PEMer	PEM-based	(Korbel et al. 2009)
VariationHunter	PEM-based	(Hormozdiari et al. 2010)
commonLAW	PEM-based	(Hormozdiari et al. 2011)
GASV	PEM-based	(Sindi et al. 2009)
Spanner	PEM-based	(Mills et al. 2011)
AGE	SR-based	(Alexej Abyzov and Gerstein 2011)
Pindel	SR-based	(Ye et al. 2009)
SLOPE	SR-based	(Abel et al. 2010)
SRiC	SR-based	(Zhang et al. 2011)
Magnolya	AS-based	(Nijkamp et al. 2012)
Cortex assembler	AS-based	(Iqbal et al. 2012)
SegSeq	RD-based	(Chiang et al. 2009)
CNV-seq	RD-based	(Xie and Tammi 2009)
RDXplorer	RD-based	(Yoon et al. 2009)
BIC-seq	RD-based	(Xi et al. 2011)
CNAseg	RD-based	(Ivakhno et al. 2010)
cn.MOPS	RD-based	(Klambauer et al. 2012)
JointSLM	RD-based	(Magi et al. 2011)
ReadDepth	RD-based	(Miller et al. 2011)
rSW-seq	RD-based	(Kim et al. 2010)
CNVnator	RD-based	(Abyzov et al. 2011)
CNVnorm	RD-based	(Gusnanto et al. 2011)
CMDS	RD-based	(Qunyuan Zhang et al. 2010)
mrCaNaVar	RD-based	(Alkan et al. 2009)
CNVeM	RD-based	(Zhanyong Wang et al. 2013c)
NovelSeq	Combinatorial approach	(Hajirasouliha et al. 2010)
HYDRA	Combinatorial approach	(Quinlan et al. 2010)
CNVer	Combinatorial approach	(Medvedev et al. 2010)
GASVPro	Combinatorial approach	(Sindi et al. 2012)
Genome STRiP	Combinatorial approach	(Handsaker et al. 2011)
SVDetect	Combinatorial approach	(Zeitouni et al. 2010)
inGAP-sv	Combinatorial approach	(Qi and Zhao 2011)
SVseq	Combinatorial approach	(Zhang and Wu 2011)

Three different strategies have been in use for detection of positive selection signatures (Figure 1.16) - increased linkage disequilibrium (LD), skew of site frequency spectrum and reduced levels of genetic variation. Several tests (summarized in Table 1.4), based on these strategies, were employed in studies looking at selection signatures in sheep (summarised in Table 1.5). Of the employed tests, LD based tests (EHH, REHH, XP-EHH, iHS and Rsb) and Fst (skew of site frequency spectrum based method) have been popular. LD based tests for detection of selection signatures are based on the assumption that the frequency of a novel mutation consequent to positive selection will increase more rapidly than that of a neutral mutation (Sabeti et al. 2002). Consequently, long LD blocks involving the mutant genes could exist in populations undergoing artificial selection since there would not be enough generations to break the LD by recombination (Slatkin 2008). Hence, a high frequency and unusually long haplotype could indicate the presence of a positive selection signature. F_{ST} is widely used to detect selection regions that differ significantly in frequency between populations (Barreiro et al. 2008) and is based on the assumption that the populations have similar effective size and are independently derived from the same ancestral population. However, in real scenarios these assumptions are often unrealistic and hence, F_{ST} values are prone to bias and false positivity (Price et al. 2010).

There is no solid conclusion about which statistic is better than others. Gouveia et al. (2017) used three tests, Fst, iHS and Rsb, to detect selection signature in three sheep breeds and found the result from the three methods to be very different. Kijas et al. (2014) did a comparison between hapFLK and Fst in sheep. hapFLK is a new method to detect selection signature, which uses frequency of haplotype instead of frequency of SNP as in case of Fst. Kijas' study found the false positive ratio of hapFLK to be lower than that of Fst.

Several studies (Table 1.5) have looked at population diversity in sheep. Unique selection signatures and the genes detected in the divergent populations could be responsible for special phenotype of each population. Fariello et al. (2014) used a modified Fst test, called FLK test and hapFLK, to detect selection signatures in 74 sheep breeds (belonging to 7 groups) around the world and found that many selection signatures were involved in hair, skin and eye color. Zhao et al. (2016) found 707, 705 and 438 regions of selection signatures using EHH based method in three breeds. Liu et al. (2016) found many unique selection signatures in three Chinese short fat-tailed sheep using Fst and Hp (average pooled heterozygosity).

Selection signature has also been widely used to detect genes associated with economic traits. Pickering et al. (2013) undertook a selection signature study using Fst between daggy Romneys and feral Arapawas and found 9 significant regions, involving 35 genes, that could be responsible for dagginess. Kim et al. (2016) found several hot and arid environment associated regions by overlapping the results of selection signatures between sheep and goat. Yuan et al. (2017) located 6 regions which were associated with tail type. Zhi et al. (2018) successfully revealed that c.G334T mutation in T/brachyury gene directly results in the short-tail phenotype in sheep, based on sequencing data. McRae et al. (2014) detected several parasite associated selection signature regions using Fst and EHH in sheep.

Finally, selection signatures located on sex chromosomes have also been reported. Zhu et al. (2015) found many selection signatures on X chromosome, which indicated that sex chromosome might be responsible for sheep's domestication.

Table 1.4 Summary of selection tests used in published studies pertaining to selection signatures in sheep

(adapted from Randhawa et al. 2016, with permission, from PLoS One, Vol.11©2016 by Public Library of Science; permission license: Creative Commons Attribution (CC BY) license)

Test	Description	Reference
Fst	Fixation Index (Population differentiations): Detects both newly arising and pre-existing variation under selection by measuring the allelic diversity between populations versus within population	(Akey et al. 2002; Nicholson et al. 2002; Weir and Cockerham 1984; Weir and Hill 2002)
ZHp	Z-transformed Heterozygosity Value: Detects selective sweeps by counting alleles in a sliding window centered on a candidate SNP, then calculates heterozygosity scores (Hp) from the pool of samples from within a population (breed) and extreme (negative) Ztransformed Hp values represent reduction in heterozygosity in the candidate regions	(Rubin et al. 2010)
EHH	Extended Haplotype Homozygosity: Detects positively selected regions carrying frequent haplotypes with unusually high long-range LD patterns within a population	(Mueller and Andreoli 2004; Sabeti et al. 2002)
REHH	Relative EHH: Detects evidence of recent selection on relatively high frequency haplotypes within a population by comparing the EHH of the tested core haplotype to that of other core haplotypes present at a locus to correct for local variation in recombination rates	(Sabeti et al. 2002)
XP-EHH	Across Population EHH: Detects selective sweeps by comparing EHH across populations in which selected alleles (at core haplotypes) have risen to near fixation in one but not all populations	(Sabeti et al. 2007)
iHS	Integrated Haplotype-homozygosity Score: Detects evidence of recent positive selection at a locus based on the differential levels of LD surrounding a positively selected (derived) allele (at intermediate frequencies) compared to the background (ancestral) allele at the same position within a population	(Voight et al. 2006)
Rsb	Across Population iES: Detects recent selection on completely or nearly fixed selective sweeps by comparing the single locus iES associated with the same site and genomic region across populations. Rsb and XP-EHH are based on similar assumptions to target haplotype decay, so they can be substituted.	(Tang et al. 2007)
FLK	An extension of Lewontin and Krakauer (LK) test that accounts for historical branching and heterogeneity of genetic drift, using a phylogenetic estimation of the population's kinship (F) matrix	(Bonhomme et al. 2010)
hapFLK	Incorporates hierarchical structure of populations, similar to FLK, but the test is extended to account for the haplotype structure in the sample, using a multipoint linkage disequilibrium model	(Fariello et al. 2013)

Table 1.5 Summary of studies pertaining to selection signatures in sheep

Reference	Platform	Assembly	Breeds (samples)	Statistic	Traits
(Pickering 2013)	Ovine 50k	Ovr v_2.0	2 breeds (83)	FST	dagginess
(Fariello et al. 2014)	Ovine 50k	Ovr v_3.1	74 breeds (3000)	hapFLK, FLK	population diversity
(Gouveia et al. 2017)	Ovine 50k	Ovr v_3.1	3 breeds (87)	Fst, his, RsB	population diversity
(Kijas 2014)	Ovine 50k	Ovr v_3.1	2 breeds (2819)	hapFLK	polled and horned
(Kim et al. 2016)	ovine 50k goat 50k	Ovr v_3.1 Caprine v2.0	2 breeds (127)	F_{ST}, iHS	hot arid environment
(Liu et al. 2016)	HiSeq2000	Ovr v_3.1	3 breeds (45)	F_{ST}, ZH_P	population diversity
(Rochus et al. 2017)	ovine 600k	Ovr v_3.1	27 breeds (691)	FLK, hapFLK	population diversity
(Yuan et al. 2017)	ovine 50k	Ovr v_3.1	6 breeds (122)	Fst, hapFLK	fat and thin tail
(Zhao et al. 2016)	ovine 50k	Ovr v_3.1	3 breeds (329)	iHS, XP-EHH	population diversity
(Zhi et al. 2018)	HiSeq2000	Ovr v_3.1	2 breeds (200)	ZH_P, Fst	short and long tail
(Zhu et al. 2015)	ovine 50k	Ovr v_3.1	3 breeds (148)	iHS, Fst	population diversity
(McRae et al. 2014)	ovine 50k	Ovr v_3.1	2 breeds (180)	Fst, iHS, XP-EHH	resistance or susceptibility to gastrointestinal nematodes

1.6 Somatic mosaicism of CNV

Somatic mosaicism is the occurrence of two genetically distinct populations of cells within an individual, derived from a postzygotic mutation. Somatic mosaicism (SM) can be either genotypic or phenotypic. Phenotypic somatic mosaicism could be caused by a the genetic mutation or other non-genetic factor, such as epigenetic, resulting in altered gene expression while the reason for genotypic somatic mosaicism is the result of mutation or recombination of DNA sequence. Detection of SM can be traced back to 1914, when abnormal karyotypes were detected in cancer tissue by Theodor Boveri (1929). After that, only a few studies were undertaken in this area until the 1970s and 1980s, when the relationship between somatic cell gene rearrangement and the functional diversity of immunoglobulin was discovered (Brack et al. 1978; Tonegawa 1983). During the past decade, advances in molecular techniques have enabled the discovery of an association of SM with cancer (Vogelstein et al. 2013), neurological disease (Poduri et al. 2013), autism (Sanders et al. 2012) or ageing (Hoeijmakers 2009; Kennedy et al. 2012). Apart from humans, SM has been reported in plants (Gill et al. 1995) and animals (Schaible 1963).

Currently, there are five major genotypic somatic mosaicism (Campbell et al. 2015): 1) chromosomal aneuploidy and large-scale structural abnormalities, 2) CNV and other structural variation, 3) single nucleotide variants (SNVs) and small insertions and deletions, 4) trinucleotide and other repeat expansion, and 5) contraction and autonomous mobile elements insertions. This thesis focused on CNV based SM.

So far, almost all CNV based SM studies were reported in humans. Žilina et al. (2015) detected five mosaic copy neutral loss of heterozygosity (CN-LOH) regions in three out of four individuals (75%), showing mosaic CNV to be a common phenomenon. O'Huallachain et al. (2012) identified 73 high-confidence mosaic CNVs, out of which seven CNVs were

common in different individuals, indicating potential hotspots might exist for somatic genomic variation. Moreover, Bruder et al. (2008) found CNVs exist within concordant and discordant phenotypical monozygotic twins (MZ) because MZ are considered genetically identical since they are derived from the same zygote. Therefore, any genotypic difference between twins means an irrefutable case of mosaicism. Finally, somatic CNV could change over time. Forsberg et al. (2012) compared the CNVs between difference ages, using a human blood cells, instead of tissues. They found the CNVs to be changing with age, which resulted in accumulation of aberrations that could finally change the phenotype. CNV based SM has beendetected in animals. Oluwole et al. (2016) found 57% of CNVs detected were different within the same cattle and the copy number of a gene, TSPY, varied significantly among several tissues.

1.7 Overall summary and thesis objectives

SNP has so far been the most popular gene marker for GWAS in animals as well as human. A relatively new marker, CNV, is being widely investigated in human, animals and plants, during the past decade. Three different platforms, aCGH, SNP microarrays and NGS, have been in use for detecting CNV. Of them, SNP microarrays have been most popular for CNV detection owing to low cost. Several algorithms have been in use for CNV detection based on SNP microarrays, however, there has been no conclusive evidence regarding the best performing one. Compared to other domestic animals, CNV has been poorly investigated in sheep, with only two published studies by the time the current project was initiated (early 2014). Hence, a study was undertaken (Chapter 2) with the following objectives.

a) Detect CNVs using 50k SNP microarray
b) Compare three different algorithms for CNV detection based on SNP microarrays
c) Compare the difference of CNV between 5 sheep breeds

NGS is another platform for CNV detection. Although relatively more expensive (compared to SNP microarrays), it provides a higher resolution for CNV detection. However, no paper has so far been published on sheep, in this area. Therefore, a separate study (Chapter 3) was undertaken with the following objectives.

a) Detect CNV in sheep using NGS data
b) Study the inheritance pattern of the detected CNVs

The extent of CNV, as well as inheritance of CNV were revealed in the first two studies (Chapters 2 and 3). Subsequently, the utility of CNV as a genetic marker for quantitative traits was tested in a subsequent study (Chapter 4). Utilising CNV detected based on high-density SNP microarrays, a GWAS was conducted on two lines of sheep selectively bred for gastrointestinal nematode resistance or resilience. The study objectives were:

a) Detect CNVs using high-density (600k) SNP microarray
b) Undertake SNP-based GWAS for three traits pertaining to gastrointestinal nematodiasis
c) Undertake CNV-based GWAS for the three traits and compare the results with those from SNP-based GWAS

Somatic mosaicism of CNV has been reported as a common phenomenon in human and other animals. However, there has been no such study reported in sheep. Therefore, somatic mosaicism of CNV between tissues was explored (Chapter 5) in sheep, adults as well as foetuses. The study objectives were:

a) Detect CNVs in adult and foetal tissues, using SNP microarray (600k)
b) Investigate between tissue CNV mosaicism in adults as well as foetuses

The final study (Chapter 6) in this project was designed to identify SNP based selection signatures in two lines of sheep selectively bred for gastrointestinal nematode resistance or resilience. Also, CNV differences between the lines were explored.

a) Detect SNP based selection signatures using within-line EHH and between line Rsb and XP-EHH tests
b) Compare the overlap of SNP-based signatures with CNV regions

Chapter 2

Genome-wide detection of autosomal copy number variants in several sheep breeds using Illumina OvineSNP50 BeadChips

Juncong Yan[a], Hugh T. Blair[a], Mingjun Liu[b], Wenrong Li[b], Sangang He[b], Lei Chen[b], Keren E. Dittmer[a], Dorian J. Garrick[c], Patrick J. Biggs[a], Venkata S.R. Dukkipati[a*]

Published in Small Ruminant Research, 2017, 155: 24-32.
(doi:10.1016/j.smallrumres.2017.08.022)

[a] IVABS, Massey University, Palmerston North 4442, New Zealand
[b] Key Laboratory of Genetics Breeding and Reproduction of Grass feeding Livestock MOA P.R. China; Key Laboratory of Animal Biotechnology of Xinjiang, Xinjiang Academy of Animal Science, Urumqi, Xinjiang, P.R. China
[c] IVABS, Massey University, Homestead, Ruakura, Hamilton 3214, New Zealand

2.1 Abstract

Characterising copy number variants (CNV) of genes has gained popularity in human and animal studies since 2006. Several studies have revealed CNVs play an important role in phenotypic diversity. However, only five papers have been published about CNV in sheep. This study detected CNV in 385 sheep belonging to different genetic groups, genotyped using Illumina OvineSNP50 BeadChips and analyzed using SVS, PennCNV, and cnvPartition algorithms. In total, SVS detected 29,935 and 33,880 CNV segments by univariate and multivariate methods, respectively, that merged into 749 CNV regions (CNVR) (size range: 15.3 to 6,600.0 kb). PennCNV and cnvPartition algorithms found 4,758 and 6,676 CNV segments, that merged into 464 (size range: 16.4 to 2,108.8 kb) and 104 (size range: 87 to 12,093.7 kb) CNVR, respectively. Most of the detected CNVRs were losses and only 69 CNVRs were detected by all three algorithms. A total of 4,635 Ensembl genes were identified in CNVR. In addition, the study revealed huge differences in CNV segments as well as CNVR among five breeds that were compared.

***Keywords*:** CNV, sheep, Illumina OvineSNP50 BeadChip, SVS, PennCNV, cnvPartition.

2.2 Introduction

Copy number variants (CNV) are defined as segments of DNA (larger than 1 kb) displaying copy number differences such as gains (insertions or duplications) or losses (deletions or null genotypes) (Feuk et al. 2006; Scherer et al. 2007). CNV in genes represent a different kind of genetic variation from single nucleotide polymorphism (SNP) and have been shown to contribute to genetic variation in production and disease traits. Initial research on CNV traces to 1936, when Bridges (1936) found an association between the BAR “gene”, duplication of a part of a chromosome, and the eye size in *Drosophila melanogaster*. With the development of molecular biology, the molecular structure of CNV began to be characterized (Iafrate et al.

2004; Sebat et al. 2004; Tuzun et al. 2005). The first-generation CNV map of the human genome comprised 1447 CNVs and was constructed by (Redon et al. 2006) using a SNP microarray.

Since 2006, several studies in humans and animals have identified CNV polymorphisms and shown associations with morphological as well as complex disease traits. It is estimated that thousands of genes (about 12% of the human genome) are variable in copy number and are likely to be responsible for a significant proportion of normal phenotypic variation (Carter 2007). The effects of CNV on phenotype could be due to changes in gene expression levels, either directly by duplication or deletion of a gene, or indirectly through position effects on downstream pathway and regulation networks (Dermitzakis and Stranger 2006; Reymond et al. 2007). CNVs have been shown to explain up to 18% of genetic variation in gene expression, having no overlap with that explained by SNP (Stranger et al. 2007). Several studies in humans revealed associations of CNV with important genetic diseases like Williams-Beuren syndrome, Angelman syndrome, breast cancer, etc. (Aitman et al. 2006; Chen et al. 2011; Fanciulli et al. 2007; Gimelli et al. 2003; Gonzalez et al. 2005; Hollox and Hoh 2014; Krepischi et al. 2012; Kurotaki et al. 2003; Morrow 2010; Osborne et al. 2001; Yang et al. 2007).

A number of studies have investigated CNV in animals, including cattle, sheep, goat, pig, horses, dogs, chicken, turkeys and ducks (Clop et al. 2012). Noteworthy significant CNV relationships identified in livestock include: a duplication of the agouti signalling protein gene, that results in white coat colour in sheep (Norris and Whan 2008) and goat (Fontanesi et al. 2009) and duplication of CIITA, a trans-activator of MHC II, associated with nematode resistance in Angus cattle (Hou et al. 2012c; Liu et al. 1994). In sheep, only five studies on CNV have so far been published: (Fontanesi et al. 2011) and (Hou et al. 2015) published

CNV maps based on array comparative genome hybridization (aCGH), (Liu et al. 2013) and (Ma et al. 2015a) detected CNV using 50k SNP microarray, while (Jenkins et al. 2016) used multiple approaches (aCGH validated by cross-comparison between SNP array and whole genome sequencing). These approaches have different advantages and disadvantages; while aCGH provides excellent performance in signal to noise ratios, SNP microarrays are cheaper and higher-throughput (Peiffer et al. 2006). Almost all studies in domestic animals, especially those investigating large populations, have used SNP microarrays for CNV detection.

In the current study, we compared three CNV calling algorithms (Golden Helix SNP variation suite or SVS, PennCNV, and cnvPartition) for detecting CNV in the sheep genome, based on 50k SNP genotype data from 385 sheep belonging to nine genetic groups. PennCNV (Wang et al. 2007) is a free programme that can process SNP data from Illumina or Affymetrix platforms, and has been the preferred algorithm for CNV calling in animal studies. It is based on a Hidden Markov Model (HMM) and utilises multiple sources of information, such as the Log R ratio (LRR), the B allele frequency (BAF), the population frequency of B allele (PFB) and the distance between neighbouring SNP, for CNV calling. In contrast, cnvPartition is a plug-in function in the Genome Studio software (Illumina, CA, USA). It applies LRR and BAF to detect CNV and can only handle Illumina SNP data. SVS is a commercial CNV caller marketed by Golden Helix, MT, USA (http://goldenhelix.com). It employs only LRR to detect CNV on either a per-sample (univariate) or multi-sample (multivariate) basis.

2.3 Materials and methods

2.3.1 Materials

The ovine 50k SNP genotypes data from 447 male and female sheep were utilised for this study. Of these, 171 animals were subjects of studies (Zhao et al. 2011; Zhao et al. 2012) on different inherited genetic conditions undertaken at the Institute of Veterinary Animal and

Biomedical Sciences, Massey University, New Zealand, while the remaining 276 animals were from a study at Xinjiang Academy of Animal Science, China. Breed details of the animals are shown in Table 2.1. DNA was extracted from either blood or fresh/formalin-fixed tissues using standard extraction protocols and genotyping performed using the Ovine SNP50 BeadChip containing 54,241 SNPs (Illumina, San Diego, CA, USA), with standard Infinium protocols. In all, the SNP genotypes from 447 animals were derived from 7 batches.

2.3.2 Quality control

The ovine SNP50 BeadChip was designed based on the genome assemblyOar_v1.0. However, the latest manifest file, containing reference to position of markers in the latest genome assembly, Oar_v3.1 (http://www.livestockgenomics.csiro.au/sheep/oar3.1.php), was employed for all analyses. Raw SNP output data (idat files) were loaded into GenomeStudio® (V2011.1, Illumina, San Diego, CA, USA) software to extract genotype calls, error rate, P50 GenCall score, signal intensity (LRR) and allelic intensity (BAF) ratios for each SNP. Samples with call rate <95% were excluded. Subsequently, LRR from GenomeStudio were inputted into SVS (v8.4.1, Golden Helix, MT, USA) using a DSF Export 4.0 plugin, in order to perform additional quality control in terms of derivative log ratio spread (DLRS) and genomic waves. The threshold for DLRS was determined by calculating the inter-quartile range (IQRs) of the distribution of DLRS and setting the outlier threshold to 1.5 IQR from the third quartile. Genomic wave detection was undertaken using a wave correction algorithm (Diskin et al., 2008) in the SVS software. In addition, samples exhibiting unusually large number of CNV segments were excluded. In total, 62 samples were excluded based on call rate, DLRS, genomic waves and unusual number of segments. The remaining 385 samples were used for detecting CNV employing three algorithms (Additional files: Table S2.1).

Table 2.1 Details of sheep that were genotyped using ovine 50k SNP microarray.

Breed	Number of animals genotyped		Number of animals included in the final analysis after quality control	
	Males	Females	Males	Females
Corriedale	8	8	6	8
Corriedale x Romney	3	12	3	12
East Friesian composite x Romney	9	11	9	11
Chinese Merino	0	276	0	251
New Zealand Merino	8	12	2	2
Romney	25	28	22	24
Texel	13	10	5	8
Wiltshire	0	6	0	6
Unknown	15	3	13	3
Total	81	366	60	325
	447		385	

2.3.3 CNV detection

Three algorithms were used to detect CNV: SVS (version 8.4.1), PennCNV (version 1.0.3) and cnvPartition. SNP data pertaining to sex chromosomes was not analysed, due to allelic imbalance in males and reported unreasonably large variable regions on the X chromosome (Gurgul et al. 2015).

The cnvPartition plugin v3.2.0 (Illumina, San Diego, CA, USA) was installed into GenomeStudio® (v2011.1), and LRR and BAF of all SNPs for 385 samples were read directly. This algorithm employs LRR and BAF to detect systematic deviation in neighbouring SNP markers, representative of distinct copy numbers. A confidence score threshold of 35 was applied and the minimum number of SNP markers per CNV segment was set to three.

The PennCNV plugin (Wang et al. 2007) was installed into GenomeStudio®, and LRR and BAF of SNP from 385 samples were exported. A file containing the population frequency of B allele (PFB) of SNP was created by using the compile_pfb.pl program in PennCNV, based on SNP data from 276 Chinese Merino sheep. The GCmodel option of PennCNV was not applied in this study because this model has not been optimised yet for non-human species (Wang K, personal communication). PennCNV integrates LRR, BAF, PFB for each SNP, and the distance between adjacent SNP, into a HMM, for detecting CNV. A minimum of three SNP per CNV segment was assumed.

LRR data pertaining to 385 quality tested samples was used to perform CNV analysis using SVS (v8.4.1, Golden Helix, MT, USA). SVS employs an optimal segmenting algorithm that can delineate CNV boundaries in the presence of mosaicism, even at a single probe level. CNVs were detected using a moving window of 20,000 SNPs, with 50 segments per window, and a minimum of three SNPs per segment. CNVs were detected on a per-sample (univariate) as well as multi-sample (multivariate) basis.

2.3.4 Derivation of CNVR and construction of CNVR map

The CNV outputs from the three algorithms were inputted into CNVRuler v1.5 software (Kim et al. 2012), in order to derive CNVR. This programme produces CNVR by merging CNVs that overlap by at least one base-pair. Derived CNVR were categorised as: 'loss' (CNVR containing deletions), 'gain' (CNVR containing duplications) and 'mixed' (CNVR containing both deletions and duplications). CNVR common between any two algorithms were determined by reciprocal overlap. Finally, the CNVR were mapped to the sheep genome using a custom written code in R (https://www.r-project.org/).

2.3.5 Gene content of CNVR and functional annotation

Gene content of CNVR was retrieved from the Ensembl Genes 82 database for Oar_v3.1 dataset, using the BioMart browser (http://www.biomart.org/). The results were then inputted into Uniprot (http://www.uniprot.org/) to get NCBI gene and protein ID. Functional annotations of identified genes were made using the DAVID Bioinformatics Resources 6.7 (https://david.ncifcrf.gov/home.jsp), to obtain gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) results.

2.3.6 CNV validation by qPCR

Selected CNVs (six) were validated by qPCR, using StepOnePlus™ Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) (L. Ma and Chung 2014). Two CNVs were selected from the results of each algorithm and DNA from eight sheep was used for validations. The *DGAT1* gene was used as reference since it has been shown to be free from copy number variation (L Fontanesi et al. 2011). Details of primer sequences, target regions in the sheep map, as well as PCR conditions are shown in Additional files: Table S2.2.

The copy number of the amplified regions was calculated by a relative standard curve method (Biosystems 2004) as described below.

$Qty = 10^{\frac{Ct-b}{m}}$, where *Qty*, m and b are the relative quantity of amplified fragment, slope and y-intercept of the standard curve.

$$copy\ \ number = \frac{Qty(\text{NormalizedTarget})}{Qty(\text{NormalizedReference})}$$

$$= \frac{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{target sample}}{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{reference sample}}$$

2.3.7 Comparison of CNV among different breeds

A comparison of CNV across five breeds was also undertaken. Since the number of Merino and Romney sheep were higher than those of the other breeds, an PCA test based on SNP data was initially undertaken within the Romney and Chinese Merino breeds by randomly dividing the number of animals into groups of around 15, to make sure that there are no obvious differences within the groups of each breed. Subsequently, CNV analyses were undertaken in five breeds: Corriedale (n=14), Chinese Merino (n=15), Romney (n=15), Texel (n=13) and Wiltshire (n =6). The CNV segments detected in these breeds by the three algorithms were overlapped using cnvRuler v1.5 to obtain CNVR for each breed. Two CNVRs sharing same areas were overlapped and considered as one CNVR. Details of genes (and their proteins) harboured in the CNVR were obtained from Uniprot database, as described in section 2.4. Also, the fixation index (F_{ST}) (Fisher 1925) between breeds was calculated based on the SNP data using SVS software.

2.4 Results

2.4.1 Genome-wide CNV detection

Illumina 50K SNP BeadChip genotypes of 385 sheep were used to detect CNV by the three algorithms (SVS, PennCNV and cnvPartition). SVS supports two methods of detection, univariate which detected 29,935 and multivariate which detected 33,880 CNV segments respectively. Among these 8,586 (28.6%) and 9,947 (29.3%) were losses, for univariate and multivariate methods respectively, 2 and 0 were gains, and the rest were neutral. The Penncnv and cnvPartition algorithms detected fewer, namely 4,758 and 6,676 segments representing 4,588 (96.4%) and 944 (14.1%) losses, and 170 and 1 gains respectively.

The loss and gain CNV segments detected by each algorithm that overlap by at least one base pair, were merged to characterize CNVR. The chromosomal distribution of CNVR detected

by the three algorithms is shown in Figure 2.1 with the two methods of SVS integrated. In total, 749 CNVRs were detected by SVS and their length ranged between 15.3 and 6,600.0 kb, with a mean and median of 189 and 118 kb, respectively, representing 748 losses and only one gain. The CNVR detected by SVS covered 141.6 Mb, representing 5.8% of the autosomes. PennCNV detected 464 CNVRs, with range, mean and median sizes of 16.4 to 2,108.8 kb, 305.5 kb and 218.1 kb, respectively, representing 426 losses, 27 gains and 11 mixed. The combined length of CNVR was 141.7Mb or about 5.8% of the total autosomal length. A total of 104 CNVRs (103 losses and 1 gain) were detected by cnvPartition and the range, mean and median of their size were 87.0 kb to 12,093.7 kb, 1,521.3 kb and 395.4 kb, respectively with a total length of CNVR being 158.2 Mb, about 6.5% of autosomes (Additional files: Table S2.3).

In order to compare the three algorithms, adjacent CNVRs that overlapped were merged into a single CNVR. After merging, there were 715, 419 and 104 CNVRs from SVS, PennCNV and cnvPartition, respectively. These included 69 CNVRs detected by all three algorithms, while 129, 13 and 13 CNVRs were mutually in common between SVS and PennCNV, SVS and cnvPartition, or PennCNV and cnvPartition, respectively. The majority of CNVRs were unique to the each of the algorithms (Figure 2.2).

Size distribution of CNVRs was analysed by dividing their length into 6 groups: 10-50 kb, 50-100 kb, 100-200 kb, 200-400 kb, 400-900 kb, and >1Mb. The frequency distribution of different size groups of CNVR (Figure 2.3) revealed obvious differences between the algorithms. The majority (77.6%) of CNVs detected by SVS were either in the 100-200 or 20-100 kb size range, while those (56.5%) found by PennCNV were either in the 200-400 or 100-200 kb range. In case of cnvPartition, 66.3% of the CNVs were either >1 Mb or were 100-200 kb long.

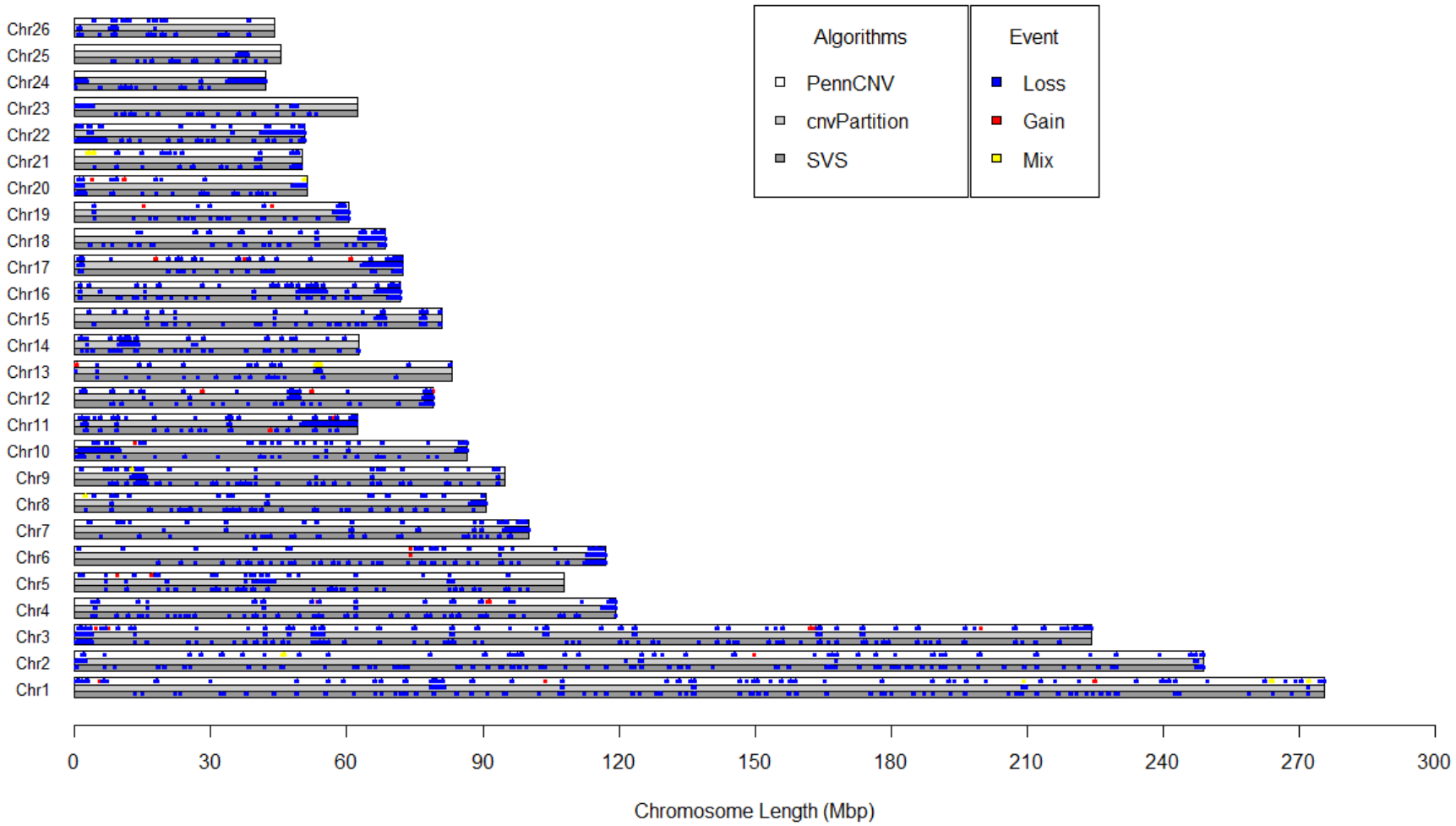


Figure 2.1 Chromosomal distribution of copy number variant regions (CNVR) detected by three algorithms.

CNVR (loss in blue, gain in red and mix in yellow) detected by PennCNV, cnvPartition and SVS algorithms are depicted across white, light grey and dark grey bar, respectively, for each chromosome.

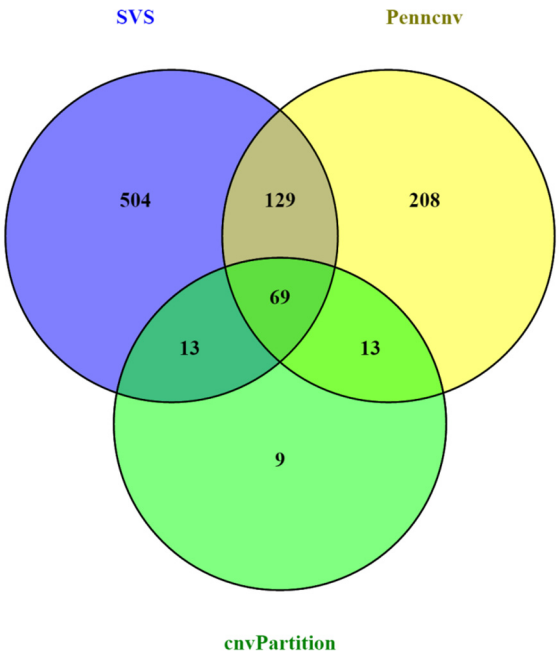


Figure 2.2 Venn plot of CNVR detected by three algorithms.

Yellow, green and blue circles represent PennCNV, cnvPartition and SVS algorithms, respectively. Numbers in overlapping regions denote the number of CNVR common to respective algorithms while those in non-overlapping regions are unique for each algorithm.

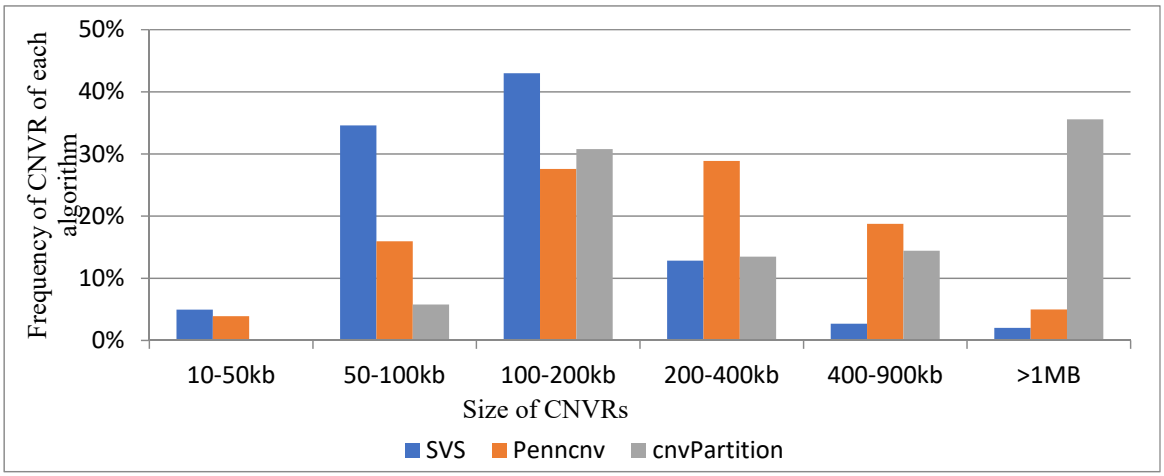


Figure 2.3 Frequency distribution of the size range of copy number variant regions (CNVR) detected by the three algorithms.

For each of the three algorithms, PennCNV (orange), cnvPartition (gray) and SVS (blue), proportional frequencies (%) of the detected CNVR in different size ranges are shown.

2.4.2 Gene content of CNVR and functional annotation of genes

Biomart probing of the Ensembl Gene 82 (http://asia.ensembl.org/index.html) database revealed the presence of 4,635 different Ensembl genes across all CNVRs from the three algorithms (Table 2.2). No genes could be retrieved from the Ensembl database for 270, 109 and 12 CNVRs detected by SVS, PennCNV and cnvPartition, respectively. The total numbers of genes found in CNVR detected by each algorithm are in Table 2.2, while details of individual genes are presented in Additional files: Table S2.4. Details of shared and unique Ensembl genes found in CNVR detected by the three algorithms are in Figure 2.4. Overall, 613 genes were found to be common for the three algorithms, while 210, 327 and 361 genes were common between SVS and PennCNV, SVS and cnvPartition, or PennCNV and cnvPartition respectively.

Since the Ensembl gene ID cannot be read directly by DAVID Bioinformatics Resources (https://david.ncifcrf.gov/home.jsp), they were inputted into Uniprot (http://www.uniprot.org/) to obtain NCBI gene and protein ID, which were then uploaded into DAVID to determine the functional significance of the identified genes (based on *Bos taurus*). Not all Ensembl genes could be matched to an NCBI Gene ID and hence, only 3529 unique NCBI gene IDs (out of the total 4635 Ensembl genes) were obtained (Additional files: Table S2.4). GO analysis indicated some GO categories are significantly ($P<0.05$) overrepresented in sheep CNVR, such as regulation of RAS GTPase activity, and lipid biosynthetic processes. KEGG analysis found some pathways to be significantly overrepresented ($P<0.05$), such as the Notch signaling pathway (details in Additional files: Table S2.5).

Table 2.2 Number of genes and proteins found in copy number variant regions (CNVR) detected by three algorithms.

Details of genes and proteins present in the detected CNVRs were obtained from Uniprot database (http://www.uniprot.org/).

Algorithm	Ensembl Genes	Uniprot	
		Mapped Genes	Proteins
SVS	2,027	1,558	1,682
PennCNV	2,068	1,583	1,710
cnvPartition	2,664	1,948	2,077
Total*	4,635	3,529	3,814

* Total across the three algorithms, after accounting for common genes and proteins

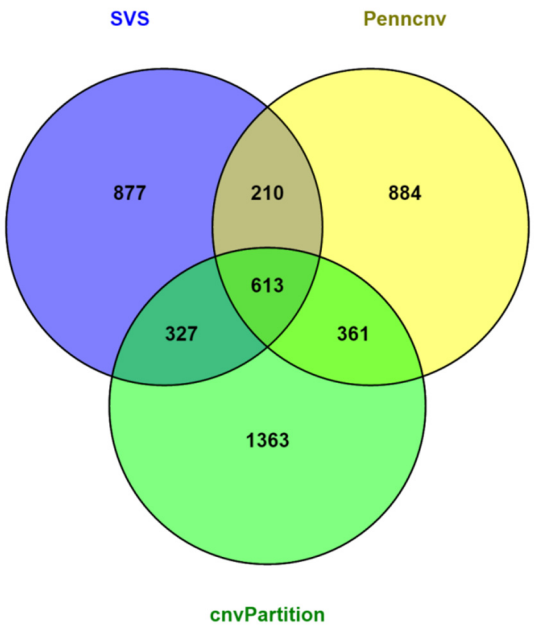


Figure 2.4 Venn plot of genes found in copy number variant regions (CNVR) detected by the three algorithms.

Yellow, green and blue circles represent PennCNV, cnvPartition and SVS algorithms, respectively. Numbers in overlapping regions denote the number of genes common to respective algorithms while those in non-overlapping regions are unique for each algorithm.

2.4.3 CNV validation by quantitative polymerase chain reaction (qPCR)

Relative qPCR was done to validate the results of the three algorithms. PCR primers were designed to amplify the DNA regions from six CNVs (two detected from each algorithm),

using DNA from eight sheep. Results showed that cnvPartition had the highest accuracy (75%) of CNV detection, followed by SVS (56%) and PennCNV (43.7%) (Table 2.3).

2.4.4 Comparison of CNVR among different breeds

CNVR among the five breeds (Corriedale, Chinese Merino, Romney, Texel and Wiltshire) are summarised in Table 2.4 and detailed results included in Additional files: Table S2.6. A Venn plot (Heberle et al. 2015)) showing huge differences in CNVR among the 5 breeds is depicted in Figure 2.5. Romney had the highest number of CNVRs (240), while only 137 CNVRs were detected in Wiltshire. However, the average number of CNVR per animal was maximum (22.8) and minimum (14.4) in Wiltshire and Corriedale, respectively. Overall, only 43 CNVRs were common in the 5 breeds. Using Ensembl (http://asia.ensembl.org/ index.html), the genes within CNVR of different breeds were detected (Table 2.4) and depicted as a Venn plot (Figure 2.6). Similar to CNVR, only 53 Ensembl genes were shared in all breeds (additional file: Table S2.6) and those pertained mostly to olfactory receptor activity and RNA binding. *Fst* between different pairs of breeds based on SNP data, ranged between 0.09 and 0.22 (Table 2.5), indicating moderate inherent genetic differences between the breeds. The PCA analysis also showed the same pattern (Figure 2.7).

Table 2.3 Results of qPCR validation of copy number variants (CNV) detected by the three algorithms.

Primers were designed for two CNVs detected by each algorithm, for validation in eight animals. Percent accuracies of CNV detection for the algorithms are presented.

Algorithms	CNV ID	Accuracy of detection	Overall accuracy for each algorithm
SVS	CNVR119	50%	56.25%
	CNVR26	62.50%	
PennCNV	CNVR9	37.50%	43.75%
	CNVR103	50.00%	
cnvPartition	CNVR5	62.50%	75.00%
	CNVR6	87.50%	

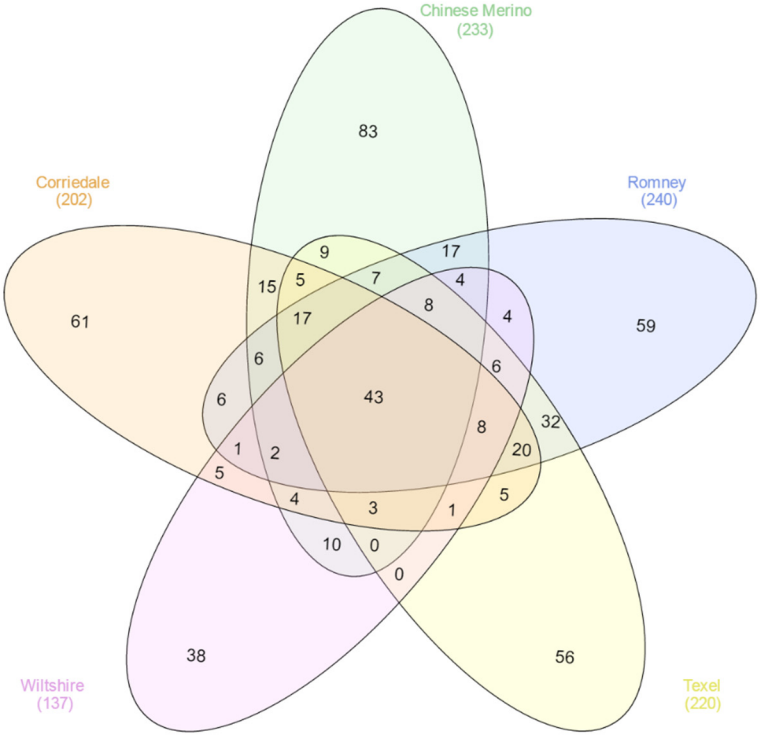


Figure 2.5 Venn plot of copy number variant regions (CNVR) detected among the five breeds of sheep.

Orange, green, blue, yellow and pink ovals represent Corriedale, Chinese Merino, Romney, Texel and Wiltshire breeds, respectively. Numbers in overlapping regions denote the number of CNVR common to respective breeds while those in non-overlapping regions are unique for each breed.

Table 2.4 Summary of copy number variant regions (CNVR) detected and their gene content in the five breeds of sheep.

CNV segments detected by the three algorithms (PennCNV, cnvPartition and SVS) were overlapped using cnvRuler v1.5, in order to obtain CNVRs in each breed, and gene content of CNVRs obtained from Uniprot database (http://www.uniprot.org/).

	Corriedale	Chinese Merino	Romney	Texel	Wiltshire
Number of animals	14	15	15	13	6
Number of CNVR	202	233	240	220	137
Number of CNVR/animal	14.4	15.5	16	16.9	22.8
Number of genes	1,069	345	1,848	1,668	585
Number of genes/CNVR	5.2	1.4	7.7	7.5	4.2

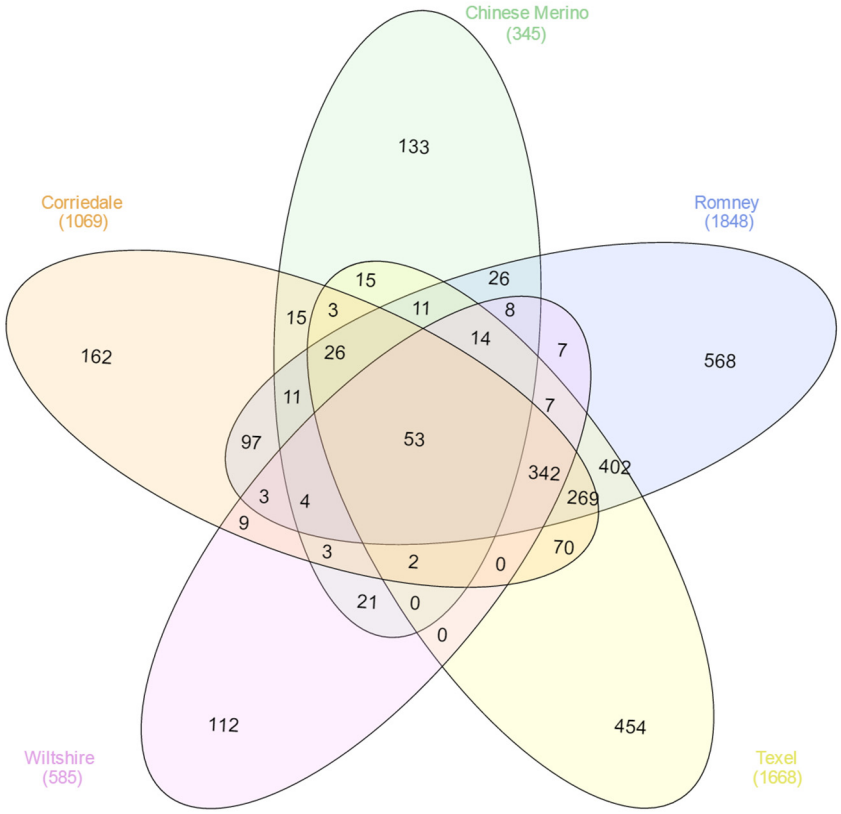


Figure 2.6 Venn plot of genes found in copy number variant regions (CNVR) detected among the five breeds.

Orange, green, blue, yellow and pink ovals represent Corriedale, Chinese Merino, Romney, Texel and Wiltshire breeds, respectively. Numbers in overlapping regions denote the number of genes common to respective breeds while those in non-overlapping regions are unique for each breed.

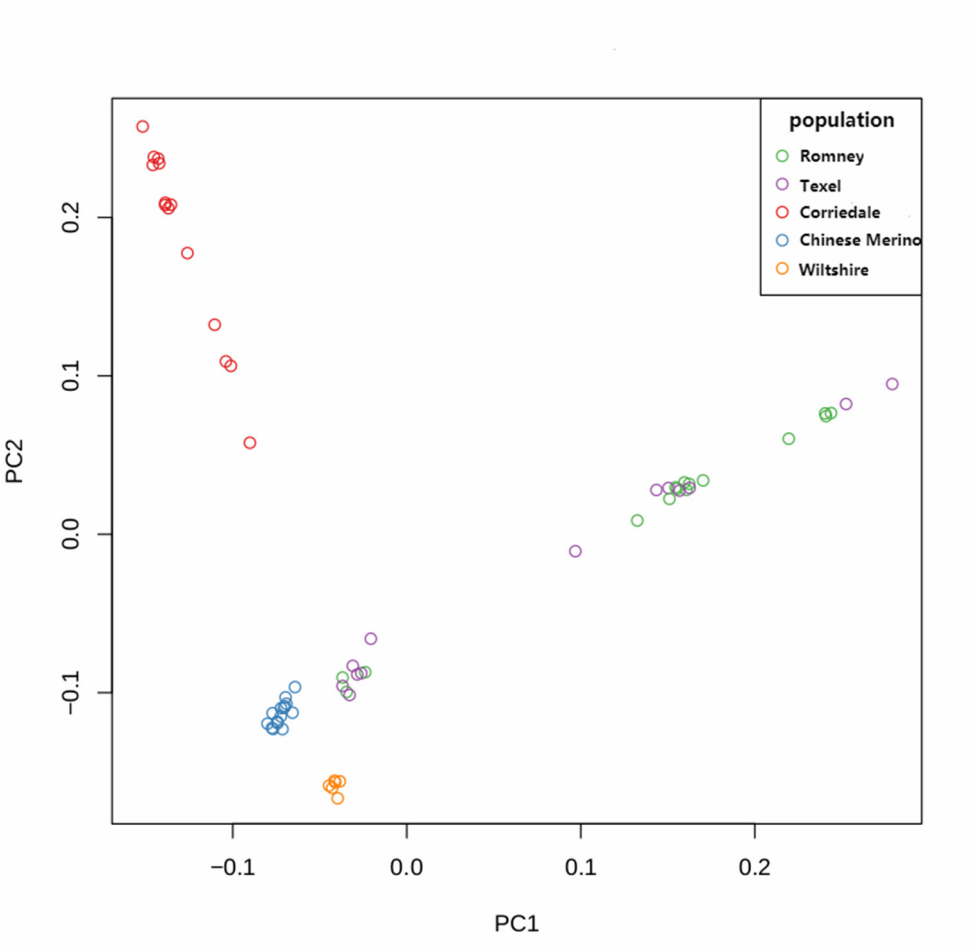


Figure 2.7 Principal component analysis plot (PC1 and PC2) showing population stratification

Green, purple, red, blue and yellow circles represent Romeny, Texel, Corriedale, Merino and Wiltshire breeds. X axis and Y axis represent PC1 and PC2, respectively.

Table 2.5 Pairwise population fixation index (F_{ST}) values for the five sheep breeds.

FST (Fisher 1925) between breeds was calculated based on SNP data using SVS software.

Breed	Corriedale	Chinese Merino	Romney	Texel	Wiltshire
Corriedale	0				
Chinese Merino	0.1278	0			
Romney	0.1827	0.1257	0		
Texel	0.1467	0.0903	0.0009	0	
Wiltshire	0.2244	0.1492	0.1913	0.1533	0

2.5 Discussion

2.5.1 Genome-wide CNV detection

In this study, three algorithms (SVS, PennCNV, cnvPartition) were employed to detect sheep CNV. Most CNV were losses, whereas gains were very rare. These findings are similar to a previous study (Xu et al. 2013) in cattle. This could be the bias because of SNP microarrays (Christensen 2010). The loss of CNV segments in an individual could result in a few genes missing, but if a gene just loses one copy, the other copy might compensate in producing the protein encoded by the gene. However, a gain of CNV might lead to the amount of gene product increasing dramatically, potentially leading to developmental malformations or alteration of function. For example, in humans, it had been revealed that duplication of a regulatory element affecting the expression of Bone morphogenetic protein 2 would result in brachydactyly type A2, a limb malformation characterized by hypoplastic middle phalanges of the second and fifth fingers (Dathe et al. 2009). However, such CNV gains might be eliminated by natural selection. In the current study, almost all the CNV losses detected by SVS (univariate 8190, 95.3% and multivariate 6988, 70.2%) and PennCNV (4554, 99.2%) were single copy losses. However, only 252 of the total 944 CNV losses (26.7%) detected by cnvPartition were single copy losses, the reaming 73.3% being double copy losses. Similar huge variation in CNV detection between different algorithms was evident in studies pertaining to cattle (Xu et al. 2013). It had been suggested in those studies that since there is no “gold standard” to estimate the accuracy of algorithms for CNV detection, it is not possible to decide which algorithm is more reliable, and hence, the best choice going forward is to use multiple algorithms to detect overlapped CNV, rather than using a single algorithm.

Comparison of the results from the three algorithms showed that SVS tended to detect relatively smaller (50-200 kb) CNVR than the other two, and this explains why more CNVR were detected by SVS than PennCNV and cnvPartition. The CNVR detected by PennCNV

showed a normal distribution of size, the majority being in the range of 100-400 kb. Most (>35%) of the CNV detected by cnvPartition were larger than >1Mb in size, and this could be why fewer CNVR were detected by that algorithm (Figure 2.3).

In this study, just 69 CNVRs (5.6% of all CNVR) were found to be common to all three algorithms and 155 CNVRs (12.5% of all CNVR) were common to the three combinations of two algorithms. The majority of CNVRs were unique to each of the algorithms, indicating it is necessary to set up a standard to evaluate algorithms for CNV detection (Liu et al. 2013; Winchester et al. 2009; Xu et al. 2013). An issue that might have partly contributed to observed differences between algorithms is the CG model. CG model that is commonly used in case of PennCNV for adjustment of the differences in GC-content (e.g. waviness) across the human genome (Diskin et al. 2008) was not employed in this study as this model is yet to be optimised for non-human species (Wang K, personal communication).

2.5.2 Gene content of CNVR and functional annotation of genes

In total, 4635 Ensembl genes were found within the detected CNVRs (Table 2.2). Because a few CNVRs overlapped between algorithms, only a small number of genes were found to be in common. Furthermore 1577 Ensembl genes were excluded from the functional enrichment analysis because DAVID could not read the Ensembl gene ID of sheep directly and this resulted in a partial loss of gene information. Eventually, 57, 32 and 40 GO terms pertaining to biological processes, cellular components and molecular function, respectively, were detected (Additional files: Table S2.5). Ten GO terms (GO:0000166, GO:0046872, GO:0043169, GO:0043167, GO:0008270, GO:0046914, GO:0003723, GO:0032553, GO:0032555 and GO:0017076) found in this study overlapped with those found in Liu et al. (2013). These GO terms pertained to binding genes such as nucleotide binding, metal ion binding and cation binding. However, there was no overlap in the detected GO terms among

the three previous studies in sheep (Fontanesi et al. 2011; Hou et al. 2015; Ma et al. 2015a). It is interesting to note that just only one KEGG PATHWAY (bta04330) passed the *P* value for multiple testing for false discovery rate (FDR), such as the Benjamini test, which indicates that there could be a few false positive results in this study.

2.5.3 CNV validation by qPCR

In this study, relative qPCR was used as a method to evaluate the accuracy of the CNV detection algorithms. Results of those validations (Table 2.3) revealed that cnvPartition (75%) had a higher accuracy than PennCNV (43.75%) and SVS (56.25%). However, since those validations were based on a limited number of PCR primers, this conclusion could be biased.

2.5.4 Comparison of CNVs among different breeds

Summary of the CNVR distribution across the five breeds (Table 2.4 and Figure 2.5) indicated a huge difference of CNVR between the genetic groups. There were only 43 CNVRs shared in all breeds. The number of CNVR detected was generally proportional to the sample size in each group, except for Wiltshire (Table 2.4). However, the total number of genes identified in each breed was not proportional to the number of CNVR detected. It was surprising to note that only 345 genes could be identified in the 233 CNVRs detected in Chinese Merinos (n=15). However, a very high number of genes were identified in CNVR of Romneys and Texels, despite having similar number of CNV as in Merinos (Table 2.4). A detailed look at the distribution of the CNVR in each breed in relation to the Oar 3.1 reference revealed that most CNVRs in Chinese Merino are located in non-coding regions of the chromosomes or in regions where reference is yet to be completed. The opposite was true with regard to the Corriedale, Romney and Texel breeds, where CNVR were mostly located in gene rich regions. Also, the majority of the CNVR (Figure 2.5) and genes (Figure 2.6) in

each genetic group were unique and only 43 CNVRs and 53 genes were common across the five breeds. This is unsurprising considering the genetic diversity of the breeds as evident from the F_{ST} results (Table 2.5).

2.5.5 Comparison of this study with previous studies

So far, only five studies have been undertaken on CNV in sheep; two studies (Fontanesi et al. 2011; Hou et al. 2015) used aCGH, two (Liu et al. 2013; Ma et al. 2015a) used SNP microarrays for CNV detection and one (Jenkins et al. 2016) used multi- faceted approach. Also, those studies were based on different genome builds. One (Ma et al. 2015a) study was based on Oar_v3.1, one on Btau_v4.0 (Fontanesi et al. 2011), two on Oar_v1.0 (Hou et al. 2015; Liu et al. 2013) and another (Jenkins et al. 2016) on UMD3_OA. Comparison of detected CNVR across the studies (Table 2.6) reveals that the number of CNVR detected (adjusted to sample size) by the first (Fontanesi et al. 2011) and second (L. Hou et al. 2015) studies using aCGH are similar. However, Jenkins et al.'s study (Jenkins et al. 2016) detected as many as 3,488 CNVRs in only 36 sheep, which could be due to the employed high density of aCGH probes (2.1 million) and the multiple approaches of CNV detection. The number of CNVR detected in the current study using PennCNV was the highest compared to that in the other two studies (Liu et al. 2013; Ma et al. 2015a) that employed the same algorithm. This indicated that the results of CNV calling using a particular algorithm could be influenced by population size, breed and the employed genome assembly version.

2.6 Conclusion

In summary, huge differences in CNVR number and size were evident between algorithms, indicating the necessity for using multiple algorithms for CNV detection. Also, there were differences between breeds.

Table 2.6 Comparison of number and size of copy number variant regions (CNVR) detected in this study with those from previous studies.

Sample Size	CNVR Number	Mean size (kb)	Median (kb)	Size range (kb)	Platform	Algorithms	Assembly	References
11	135	77.6	55.9	24.6-505	Bovine 385k aCGH		Btau_v4.0	(Fontanesi et al. 2011)
329	238	253.57	186.92	13.66-1,300	OvineSNP50K	PennCNV	Oar_v1.0	(Liu et al. 2013)
160	111	123.84	100.53	Unknown	OvineSNP50K	PennCNV	Oar_v3.1	(Ma et al. 2015a)
5	51	304.86		52-2,000	1.4 M aCGH		OaiAri1	(Hou et al. 2015)
36	3,488	19		1-3,600	2.1M aCGH		UMD3_OA	(Jenkins et al. 2016)
385	749	189	118	15.3-6,600	OvineSNP50K	SVS	Oar_v3.1	This study
	464	305.5	218.1	11.4-2,108.8		PennCNV		
	104	1,521.3	395.4	87-12,093.7		cnvPartition		

2.7 Authors' contributions

JY carried out the data analysis and drafted the manuscript, VSRD and HTB participated in the study design and manuscript preparation. ML, WL, SH, LC, KED, DJG generated raw SNP data. All authors contributed to editing the article and approved the final manuscript.

2.8 Acknowledgements

This work was supported by Massey University, International Scientific Cooperation Grant 2014DFA30970 from Ministry of Science and Technology, China and
by a subcontract of grant 2013AA102506 from China national science and technology plan. Also, the primary author had been supported by Massey University Doctoral Scholarship.

2.9 Additional files

Table S 2.1 Sample information

Table S 2.2 Details of qPCR primers and reaction conditions

Table S 2.3 Details of CNV and CNVR detected by the three algorithms.

Table S 2.4 Gene annotations details

Table S 2.5 Go Ontology and KEGG analysis results

Table S 2.6 Details of CNVR in the five breeds

Chapter 3

Detection of Copy Number Variation in sheep by whole genome sequencing

Juncong Yan[1], Hugh T. Blair[1], Keren E. Dittmer[1], Patrick J. Biggs[1], Venkata S.R. Dukkipati[1*]

To be submitted to BMC Genomics

[1] IVABS, Massey University, Palmerston North 4442, New Zealand
[*] Correspondence: R.Dukkipati@massey.ac.nz

3.1 Abstract

3.1.1 Background

Sheep (*Ovis aries*) is one of the most important livestock species in the world. Copy number variation (CNV) has become an important tool to explore genetic variation and its association with economic traits in animals. However, previous CNV studies in sheep were based on either single nucleotide polymorphism (SNP) microarrays or comparative genome hybridization arrays (aCGH). Both platforms have some natural drawbacks for CNV detection. Therefore, a study was undertaken to detect CNVs in sheep using next-generation sequencing (NGS) data.

3.1.2 Results

The average read depth of coverage of this study was from 8.6x to 12.7x and coverage to the reference genome was from 81% to 84%. In total, 1836 copy number variation regions (CNVRs) were identified from the genome of 5 Romney sheep. Of them, 1653 were losses, 181 were gains and 2 were mixes. The size of CNVRs ranged between 999 and 73,499 bp, with a mean and median of 3835 and 1999 bp, respectively. The result is much smaller than previous studies based other platforms, which ranged between 19 kb and 305.5 kb. There were 587 Ensembl genes located within the identified CNVR. Only one CNVR from this study was also detected in previous studies that used 50K and 600K SNP microarrays. Besides, the average number of CNVR per sample detected using NGS (1,836 CNVRs detected in 5 animals; this study) was much higher than that detected using either 50K SNP microarray (575 CNVRs in 545 animals) or 600 K SNP microarray (339 CNVRs detected in 93 animals), in our previous studies. The pedigree comparison showed that 355 CNVs (71%) and 360 CNVs (65.9%) of two offspring could be traced from their parents.

3.1.3 Conclusion

This study successfully detected 1,836 CNVRs, ranging between 999 and 73,499 bp, in five animals using NGS data. Compared with previous studies that used SNP arrays, NGS

supports higher resolution and detection rate. Also, a good reference genome and a high sequencing depth are essential for more efficient CNV detection using NGS.

3.1.4 Keywords

Copy number variation, sheep, whole gene sequencing, NGS, CNV

3.2 Introduction

A copy number variant (CNV) is defined as a DNA segment which is larger than 1 kb and shows differences of copy number such as gains (insertions or duplications) or losses (deletions or null genotypes) (Feuk et al. 2006; Scherer et al. 2007). This kind of genetic variation could be used as a genetic marker and associations between production and disease traits and CNVs have been discovered. The first research of CNV can be traced back to 1936, when Calvin Bridges found an association between the BAR "gene" and the eye size in drosophila (Bridges 1936). The BAR gene is a duplication of a part of a chromosome. Seventy years later, with the help of molecular biology, the structure of CNV at the molecular level began to be unveiled (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005). Redon et al. (2006) constructed the first CNV map of the human genome based on a single nucleotide polymorphism (SNP) microarray (500K) and found 1,447 copy number variable regions (CNVRs).

Since 2000, several studies of CNV polymorphisms in humans and animals have identified CNV associations with economic and disease traits. These CNV regions contained hundreds of genes, disease loci, functional elements and segmental duplications. About 12% of the human genome (~360 megabases in total) is variable in copy number (Redon et al. 2006). The CNV could influence phenotype by changing the levels of gene expression as a duplication or deletion of a gene can either directly or indirectly influence downstream pathway and regulation networks (Dermitzakis and Stranger 2006; Reymond et al. 2007). CNV is responsible for 18% of genetic variation in gene expression which cannot be

explained by SNP (Stranger et al. 2007) and is associated with many human genetic diseases (Aitman et al. 2006; Fanciulli et al. 2007; Gimelli et al. 2003; Gonzalez et al. 2005; Hollox and Hoh 2014; Kurotaki et al. 2003; Osborne et al. 2001; Yang et al. 2007).

Several studies have investigated CNV in animals, such as cattle, sheep, goat, pig, horse, dog, chicken, turkey and duck, and significant relationships between CNV and traits were identified (Clop et al. 2012). For example, a duplication of the agouti signalling protein is responsible for white coat colour in sheep (Norris and Whan 2008) and goats (Fontanesi et al. 2009) and duplication of *CIITA*, a trans-activator of MHC II, is associated with resistance to nematodes in Angus cattle (Hou et al. 2012c; Liu et al. 1994).

In sheep, several studies have been published: Fontanesi et al. (2011) made the first comparative map of CNV in the sheep genome and Hou et al. (2015) published a genome-wide analysis of CNV using a comparative genome hybridization (aCGH) method for CNV detection. Liu et al. (2013), Ma et al. (2015a) and Yan et al (2017b) detected CNVs using a 50k SNP microarray, while Jenkins et al. (2016) employed multiple methods (aCGH, SNP microarray and whole genome sequencing).

Each approach, microarray, aCGH and sequencing has advantages and disadvantages for CNV calling. While aCGH has good performance in signal to noise ratios, it is hard to detect small CNVs. SNP microarrays are cheaper and have higher-throughput (Peiffer et al. 2006), making them a good choice for large population investigations. However, SNPs are not evenly distributed across the genome and CNV located in a no–SNP region cannot be detected. Recently, Next Generation Sequencing (NGS), has been used for CNV detection. It can facilitate detection of CNV that are smaller than 10 kb (Bentley et al. 2008; Yoon et al. 2009). NGS has been successfully used to detect CNV in horses (Doan et al. 2012a) and chickens (Yi et al. 2014). The objective of this study was to examine the suitability of NGS

technology for CNV calling in sheep and to compare the detected CNV with those found in previous studies.

3.3 Materials and Methods

3.3.1 Sample collection and sequencing

Heparinised blood samples were collected from five New Zealand Romney sheep (Table 3.1) and DNA extracted by MagAttract HMW genomic DNA extraction kit (Qiagen GmbH, Germany). Resulting genomic DNA was cleaned-up using the Genomic DNA Clean and Concentrate-10 kit (Zymo Research Corp., Irvine CA, USA). DNA concentration was determined using the Qubit dsDNA BR assay kit (Life Technologies Corp., Irvine, CA, USA) with the Qubit 2.0 fluorometer (Life Technologies), and purity was determined using the Nanodrop spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, USA). The resulting high quality, high molecular weight DNA, a minimum of 5 µg at greater than 20 ng/µL, was sequenced at New Zealand Genomics Ltd (NZGL, University of Otago, Dunedin, New Zealand). NZGL prepared a fragment library using the Illumina TruSeq DNA PCR free library preparation kit (Illumina, San Diego Ca, USA) with a 550 bp insert size. Two lanes of paired-end reads (2 x 100 bp) were then obtained using an Illumina HiSeq 2000 machine (Illumina, San Diego Ca, USA). Animal 828-05-1 (male) is the offspring between 828-05-5 (ram) and 828-05-4 (ewe), while animal 828-05-3 (male) is the offspring between 828-05-5 (ram) and 828-05-2 (ewe).

3.3.2 Data preparation

The original sequencing data, containing the read information, were obtained from NZGL. Illumina NGS applies paired-end sequencing technology which tests DNA sequence from two ends so that it doubles the size of reads. In this study, two lines were used for sequencing and produced three files of each direction. The command, ‘cat’, was used to merge files having the same direction from two lines into a single file. Finally, two ‘fastq’ files, forward

and reverse, for each animal were created for further analysis. In order to ascertain the position of these reads in the genome, it was necessary to align them to a reference genome to create SAM files (Appendix 3.1).

Table 3.1 Identification and sex of Romney sheep and summary NGS data

Sample ID	Gender	Depth	Coverage (%)
828-05-1	M	8.6X	84
828-05-2	F	10.1X	81
828-05-3	M	8.9X	83
828-05-4	F	10.4X	84
828-05-5	M	12.7X	84

3.3.3 CNV calling, derivation of CNV region (CNVR) and construction of CNVR map

CNV were detected using the software CNVnator_v0.3.2 (Abyzov et al. 2011), which is a read-depth method based CNV detection program. Based on the depth of coverage (8.6X to 12.7X) in this study, a bin size of 500 bp was applied as per the CNVnator instructions. The study was confined to autosomes. Based on the suggestion from the author of CNVnator (Abyzov, personal communication), the CNVs that overlapped to the gaps of the reference genome and with *q0* (the distribution of the fraction of *q0* reads. *q0* reads are a pair of reads, of which, one can map to two or more locations, and the other is randomly chosen) ≥0.5 were excluded as quality control. The CNV outputs from CNVnator were inputted into CNVRuler v1.5 software (Kim et al. 2012), was used to derive CNVR. This programme produces CNVR by merging CNV that overlap by at least one base-pair. Derived CNVR were categorised as: 'loss' (CNVR containing deletions), 'gain' (CNVR containing duplications) and 'mixed' (CNVR containing both deletions and duplications). Finally, the CNVR were mapped to the sheep genome using a custom written code in R (https://www.r-project.org/) (Appendix 3.2).

3.3.4 qPCR validation

Two selected CNVs in an individual were validated by qPCR, using StepOnePlus™ Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) (Ma and Chung 2014). The *DGAT1* gene was used as reference since it was shown to be free from copy number variation (Fontanesi et al. 2011). Details of primer sequences, target regions in the sheep map, as well as PCR conditions are shown in Additional files: Table S3.2 qPCRresult.

The copy number of the amplified regions was calculated by a relative standard curve method (Biosystems 2004) as follow:

$$copy\ \ number = \frac{Qty(\text{NormalizedTarget})}{Qty(\text{NormalizedReference})} = \frac{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{target sample}}{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{reference sample}}$$

$Qty = 10^{\frac{Ct-b}{m}}$, where *Qty*, *Ct*, m and b are the relative quantity of amplified fragment, threshold cycle, slope and y-intercept of the standard curve.

However, it is difficult to find a standard sample as a reference which has copy number variation. Therefore, firstly, the reference was selected randomly. The copy number of reference was assumed as 1 copy, then calculate the accuracy of qPCR. After that, the copy number of reference was assumed as 2, and 3 and did the same process again. Because the gene has more than 4 copies is rare, no more assumption was set up. By comparing the accuracy between the copy numbers 1, 2, 3 the copy number which has highest correction rate is considered as the correct copy number. Of course, this method could have bias since the assumption of copy number of reference could be wrong. The copy number evaluation thresholds table is given below (Table 3.2). Based on hypothetical copy number, using the copy number value calculated by above equation of each sample, the actual copy number of each sample can be evaluated.

3.3.5 Gene annotation

By using Ensembl genome browser 85, Oar_v3.1 (http://www.ensembl.org/Ovis_aries/Info/Index), genes located in CNVRs (additional file: Table S3.1) were obtained. After converting Ensembl genes into NCBI Gene IDs by UniProt (http://www.uniprot.org/), Gene Ontology (GO) and Pathway analysis (enrichment analysis) was undertaken using DAVID_v6.8 (Beta) (https://david-d.ncifcrf.gov/home.jsp) based on Fisher's exact test or hypergeometric distribution Equation (Fisher 1937). The Bonferroni corrected P value was used to account for family-wide false discovery rate ($P < 0.05$), since Fisher's exact test was done multiple times based on the background genes.

Table 3.2 Hypothetical copy numbers of the reference and their thresholds (based on qPCR) for copy number evaluation

Hypothetical copy number of the reference sample	1 copy	2 copies	3 copies	4 copies
1 copy	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5
2 copies	0.25-0.75	0.75-1.25	1.25-1.75	1.75-2.25
3 copies	0-0.459	0.459-0.825	0.825-1.165	1.165-1.495

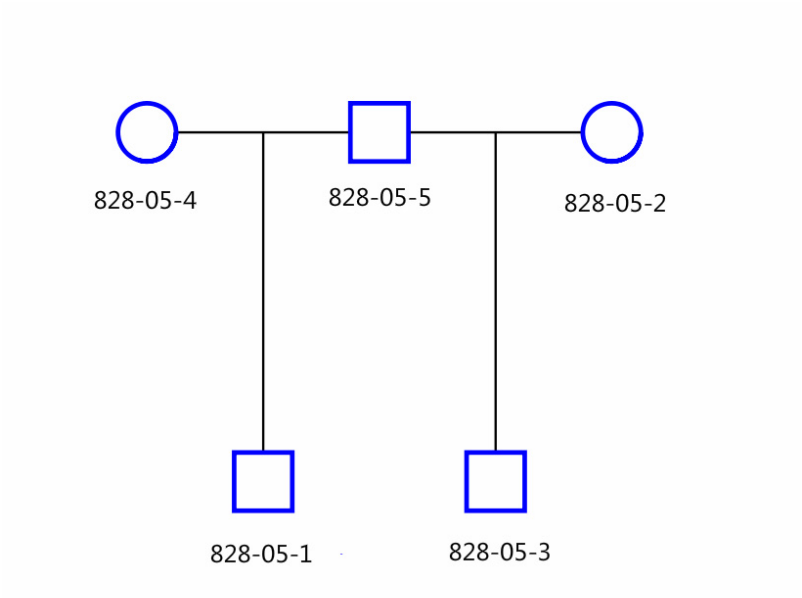


Figure 3.1 Pedigree of five sheep that were subject of the study.

Circles represent females and squares represent males.

3.3.6 Pedigree comparison

Animal 828-05-5 was the sire, while 828-05-4 and 828-05-2 were dams. Individuals 828-05-1 and 828-05-3 were lambs. The pedigree tree was used to show the relationship between them (Figure 3.1). All CNVs of each sample were sorted by genomic position to find overlapping areas. The CNVs found in parents as well as offspring were considered as inherited CNVs.

3.4 Results

3.4.1 Mapping statistics and CNV detection

The five DNA samples from New Zealand Romney sheep that were sequenced by an Illumina HiSeq 2000 machine each produced about 20 giga bases of high quality sequence data. The average read depth of coverage was from 8.6 x to 12.7 x and coverage to the reference genome was from 81% to 84% (Table 3.3). Using a custom written python script (Appendix 3.4), the genome was divided into 10 kb bins and the distribution of the sequencing depth coverage in different size bins was calculated (Figure 3.2). The bins whose read depth was less than or equal to 6 occupied 0.11% to 6.29% of the bins amongst the five individuals, while the bins whose depth was between 7 and 17 occupied 93.34% to 99.25% of the bins (Additional file Table 3.1 sheet: Depth). A bar chart (Figure 3.2) and a violin plot (Figure 3.3) were created using Excel and R (Appendix 3.5) to show the distribution of depth in each sample. As seen in those figures, individual, 828-05-5 had the highest depth.

The average number of CNV segments detected, after quality control, was 662 per sample (Table 3.3). After merging the CNV segments, 1,836 CNVRs were obtained. Of them, 1653 were losses, 181 gains and 2 were mixes (Figure 3.4). The size of CNVRs ranged between 999 and 73,499 bp, with a mean and median of 3,835 and 1,999 bp, respectively. Figure 3.5, prepared using R (Appendix 3.3), depicts the relationship between sequencing depth and number of CNVs detected in the five individuals, in 50 kb bins across the chromosomal

region, ch13:46100000-5110000. It revealed that most losses were confined to low depth areas.

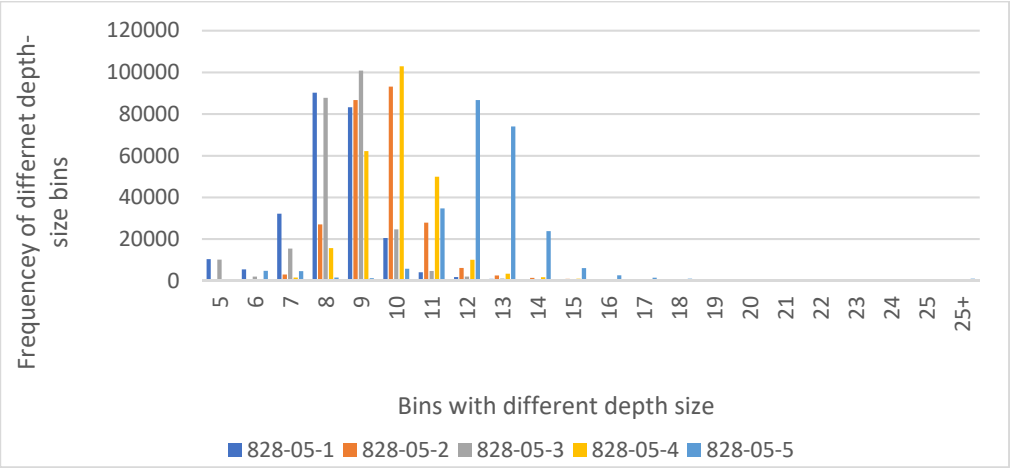


Figure 3.2 Distribution of sequencing depth-size (50 kb) bins.

The Y axis represents bins with different sequencing depth size, while X axis represents frequencies of corresponding depth-size bins. The five colours represent the five different individuals.

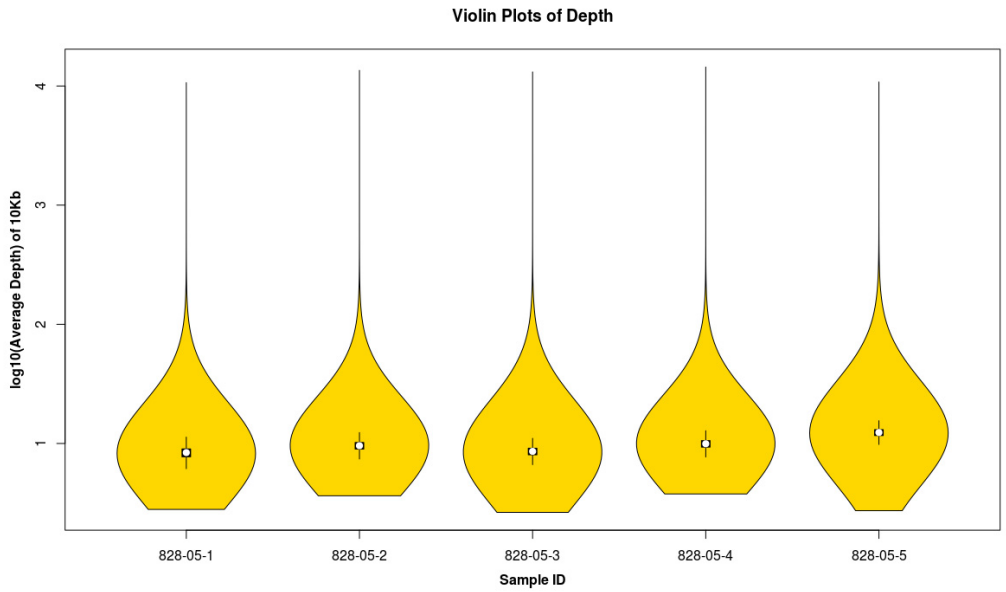


Figure 3.3 Violin plots of sequencing depth (at whole genome level) in five individuals

The X and Y axes represent the samples and the log10 (average depth of each bins), respectively.

Comparison of CNVs between the five individuals revealed that only 75 CNVs (4%) were common to all five animals and 57% (1,046 out of total 1,838) of the CNVs were unique to an animal (Figure 3.6). The majority of the detected CNVR were less than 5 kb in size (Figure 3.7).

Table 3.3 Summary of the copy number variants (CNVs) detected in five Romney sheep

Sample ID	Depth	Number of CNVs detected	Number of CNVs after quality control				
			Deletions	Duplications	Total	Mean	Median
828-05-1	8.6x	16,665	425	76	501	4,756.4	2,500
828-05-2	10.1x	18,080	514	82	596	4,096.4	2,000
828-05-3	8.9x	17,059	481	68	549	4,359.7	2,500
828-05-4	10.4x	18,392	587	89	676	4,222.6	2,500
828-05-5	12.7x	20,553	903	85	988	3,427.1	2,000
Average/sample	10.14x	18,149.8	582	80	662	3,835	1,999

3.4.2 qPCR validation

Two randomly selected CNVs detected in an individual were validated by qPCR. The observed (based on qPCR) copy numbers for the two tested CNVs matched (100%) with the predicted copy numbers (Additional file: S3.2 qPCRresult).

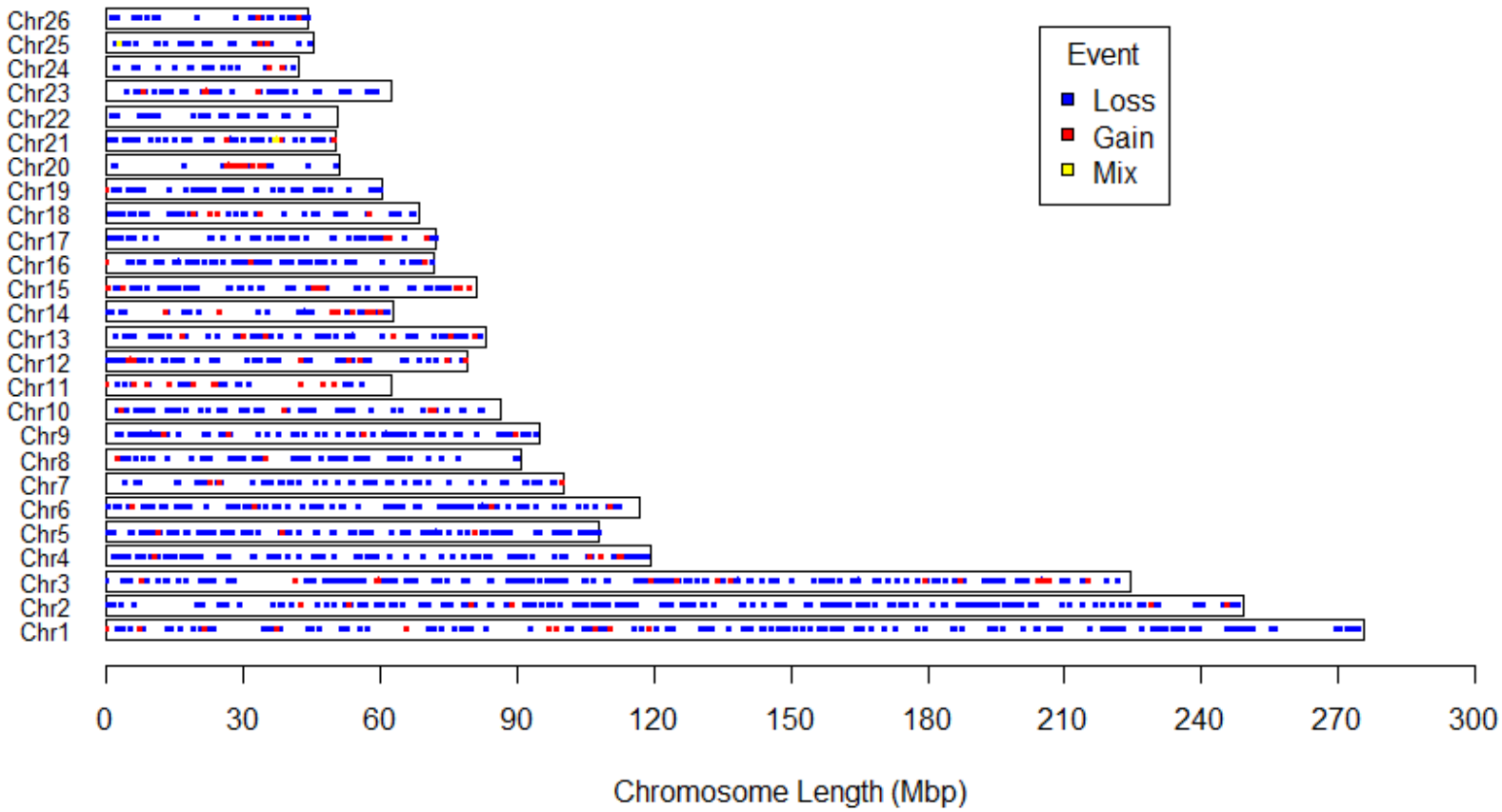


Figure 3.4 Chromosomal distribution of copy number variant regions (CNVR) detected in five Romney sheep, using whole genome sequencing data.

CNVRs (losses in blue, gains in red and mixes in yellow) are depicted across the bar for each chromosome.

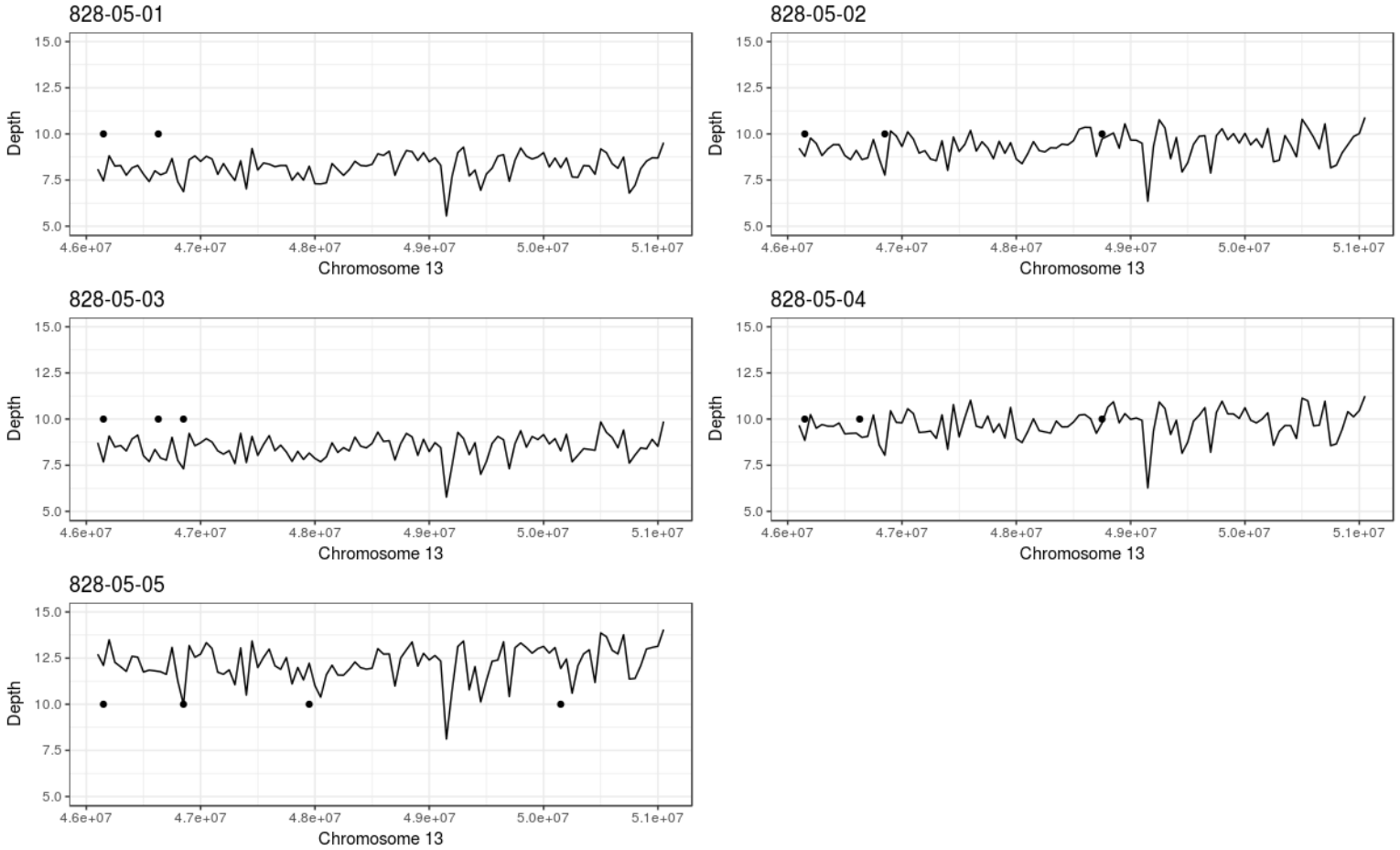


Figure 3.5 Plot showing relationship between sequencing depth and number of CNVs detected in the five individuals, in 50 kb bins across the chromosomal region, ch13:46100000-5110000.

X and Y axis in each graph represent position of chromosome and sequencing depth, respectively. The black points represent the CNVs (losses). The plots were made based on the average depth in 50 kb bins and created using an R script (Appendix 3.3). Majority of the losses were detected in low depth zones.

3.4.3 Gene annotation

In total, 587 Ensembl genes were found to be located in the detected CNVRs (additional file: Table S3.1) and NCBI gene IDs could be identified for 501 genes. GO and pathway analysis of the NCBI genes revealed that 19 GO BP (biological process), 14 GO CC (cellular component), 10 GO MF (molecular function) and 4 KEGG pathways were over-represented (P<0.05) in the identified CNVRs (additional file, Table S1). However, none of the over-represented GO categories or pathways passed multiple testing correction (Bonferroni corrected P<0.05).

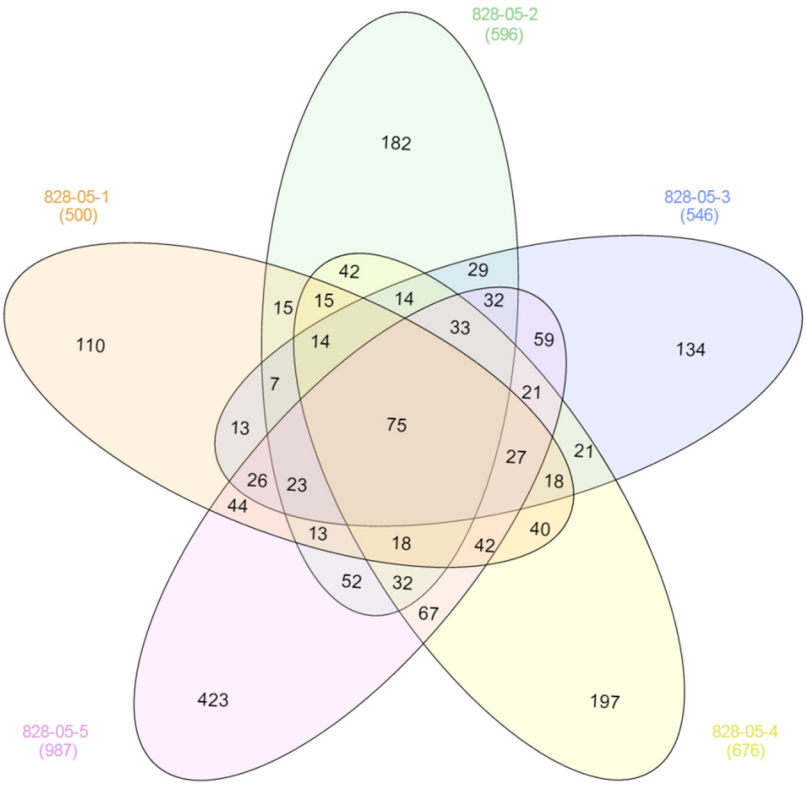


Figure 3.6 CNV comparison between five Romney sheep.

Individual sheep are shown in different coloured ovals. Numbers in overlapping regions denote the number of CNVs common to respective individuals while those in non-overlapping regions are unique for each individual. There is a slight discrepancy with regard to the number of CNV (compared those in Table 3.3) in individuals, as one large CNV detected in an individual could have been detected as several small CNVs in another individual during the analysis.

3.4.4 Pedigree comparison

In the offspring 828-05-01, 355 CNVs (71% of the total in the individual) could be traced from its parents (Figure 3.8), while in another progeny, 828-05-03, 360 CNVs (65.9% of the total in the individual) were traced from its parents (Figure 3.9). Further, out of the CNVs inherited by the two progenies (106 and 133 CNVs, respectively, by 828-05-01 and 828-05-03), exclusively from the sire, 26 CNVs overlapped (Figure 3.10).

3.5 Discussion

3.5.1 Mapping statistics and CNV detection

Figure 3.2 and 3.3 showed that the most predominant depth of the reads in the five samples was about 9X. As expected, the number of CNV detected in the samples increased with increased sequencing depth. The individual with the highest depth, 828-05-5, was found to have maximum number of CNVs while the one with the lowest depth, 828-05-1, had the least CNVs, showing that the depth of coverage is a key factor in CNV detection from NGS data. Before quality control, there were about 18,000 CNVs detected in each sample. However, because of too many gaps (about 120,000) existing in the Oar_v3.1 assembly, the number of CNVs were dramatically reduced to about 500 in each sample. These gaps are the zones on the reference genome that include highly repetitive sequences. The sequencing reads from those regions could not be mapped to the reference genome. Normally, CNVs detected around gaps are considered unreliable. A cattle study with 20 cattles (Dolezal et al. 2014) using the same software, CNVnator, identified 29,975 deletions, 1,489 duplications and 365 complex CNVRs, which were much higher than those detected in this study. However, Dolezal et al (2014) used a 20X depth and only 63,000 gaps were reported in the bovine genome assembly. Hence, a completed genome assembly and a higher depth of sequencing might be necessary for CNV detection.

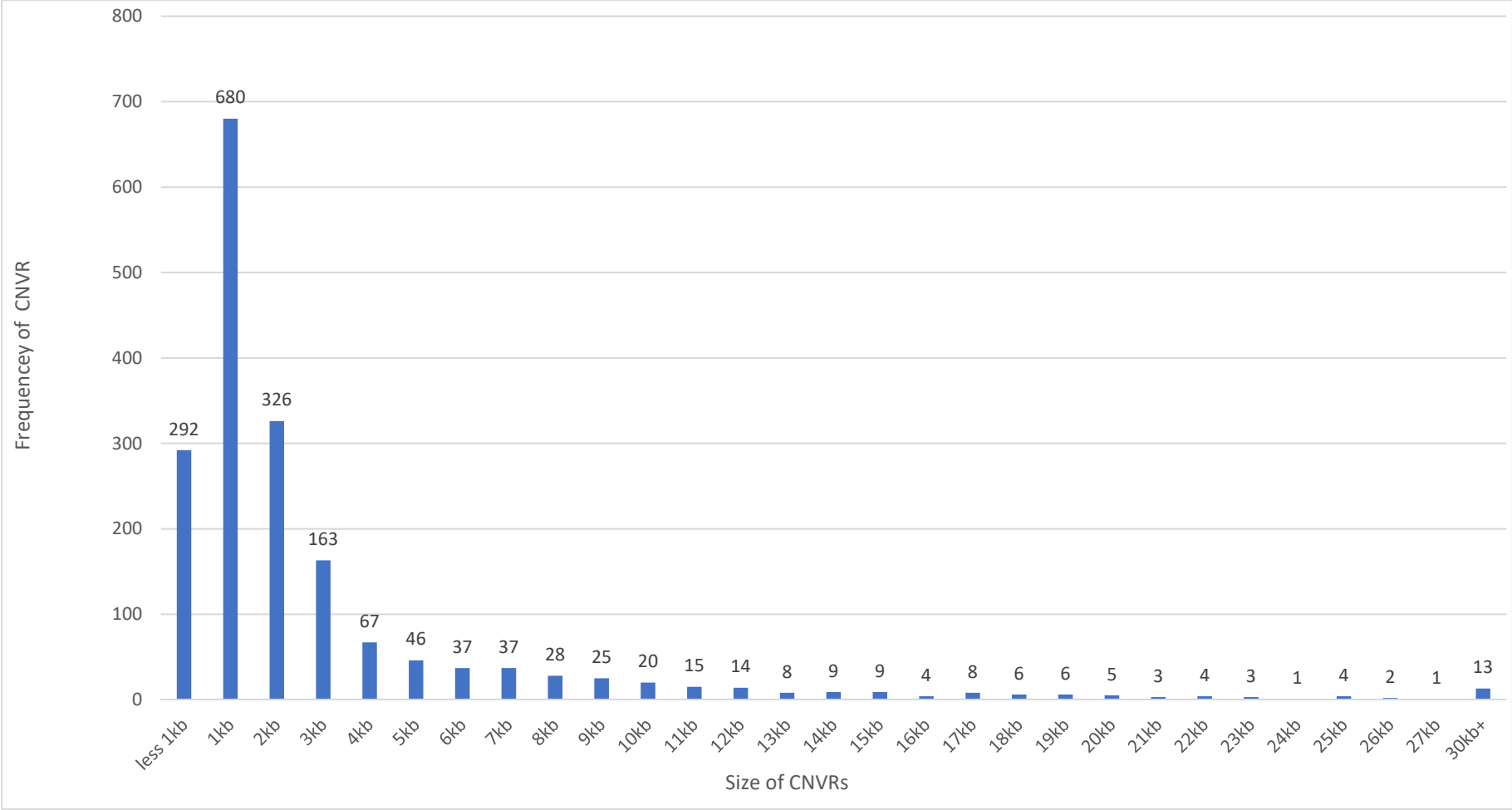


Figure 3.7 Frequency distribution of the size range of copy number variant regions (CNVR) detected in five Romney sheep, using NGS.

Frequencies of the detected CNVR in different size ranges are shown.

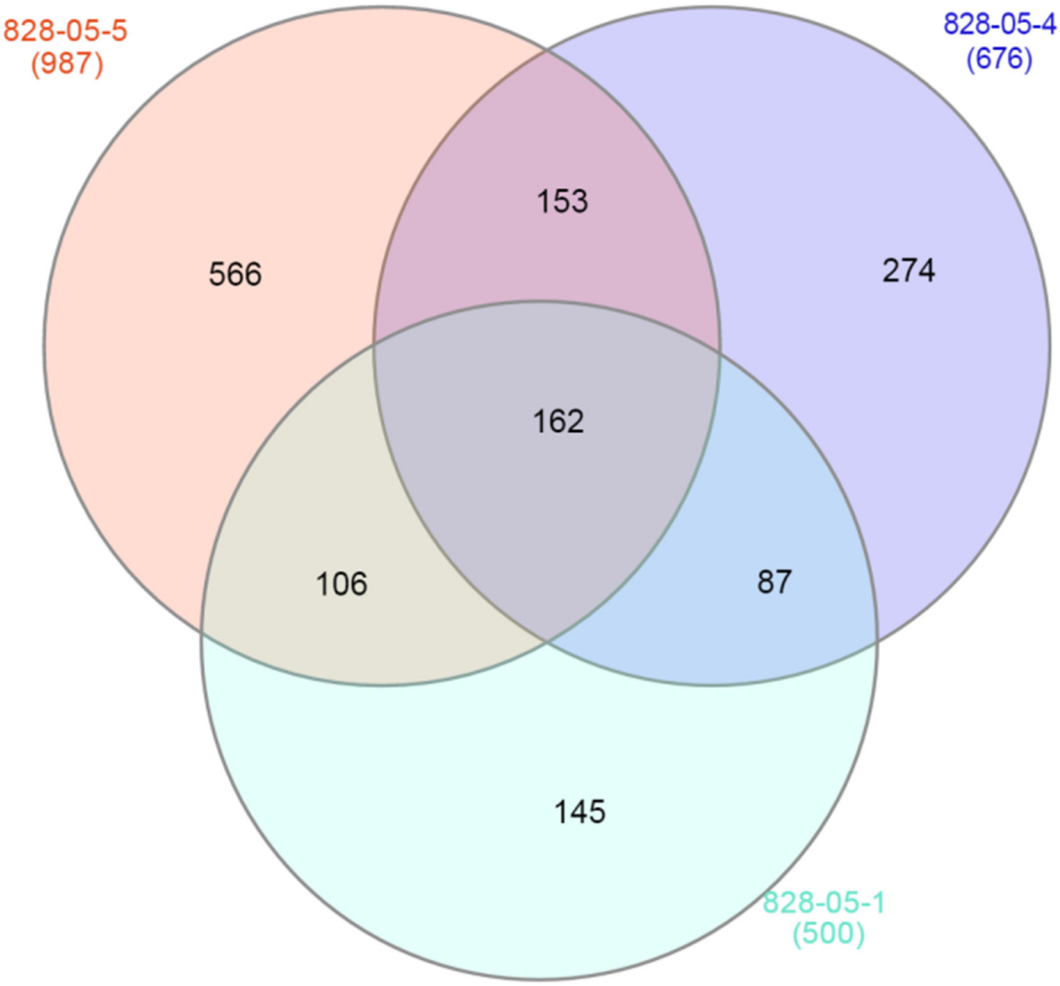


Figure 3.8 Inheritance of CNV in individual 828-05-1.

Pink, purple and blue circles represent CNVs detected in animals 828-05-5, 828-05-4 and 828-05-1, respectively. Numbers in overlapping regions denote the number of CNVs common to respective individuals while those in non-overlapping regions are unique for each individual. There is a slight discrepancy with regard to the number of CNV (compared those in sheep) in individuals as some CNVs were merged during the analysis because one large CNV could be divided into several small CNVs in another individual.

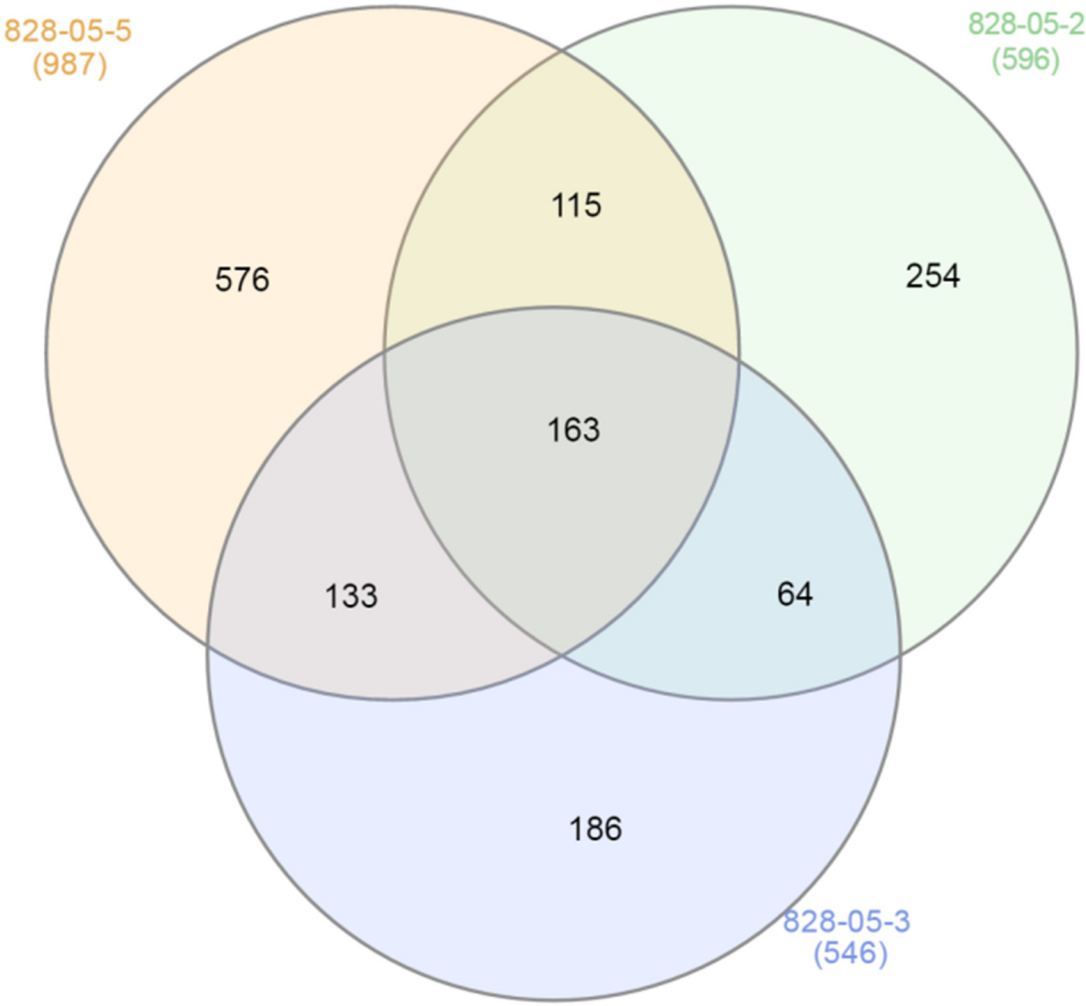


Figure 3.9 Inheritance of CNV in individual 828-05-3.

Orange, green and blue circles represent CNVs detected in animals 828-05-5, 828-05-2 and 828-05-3, respectively. Numbers in overlapping regions denote the number of CNVs common to respective individuals while those in non-overlapping regions are unique for each individual. There is a slight discrepancy with regard to the number of CNV (compared those in sheep) in individuals, as one large CNV detected in an individual could have been detected as several small CNVs in another individual during the analysis.

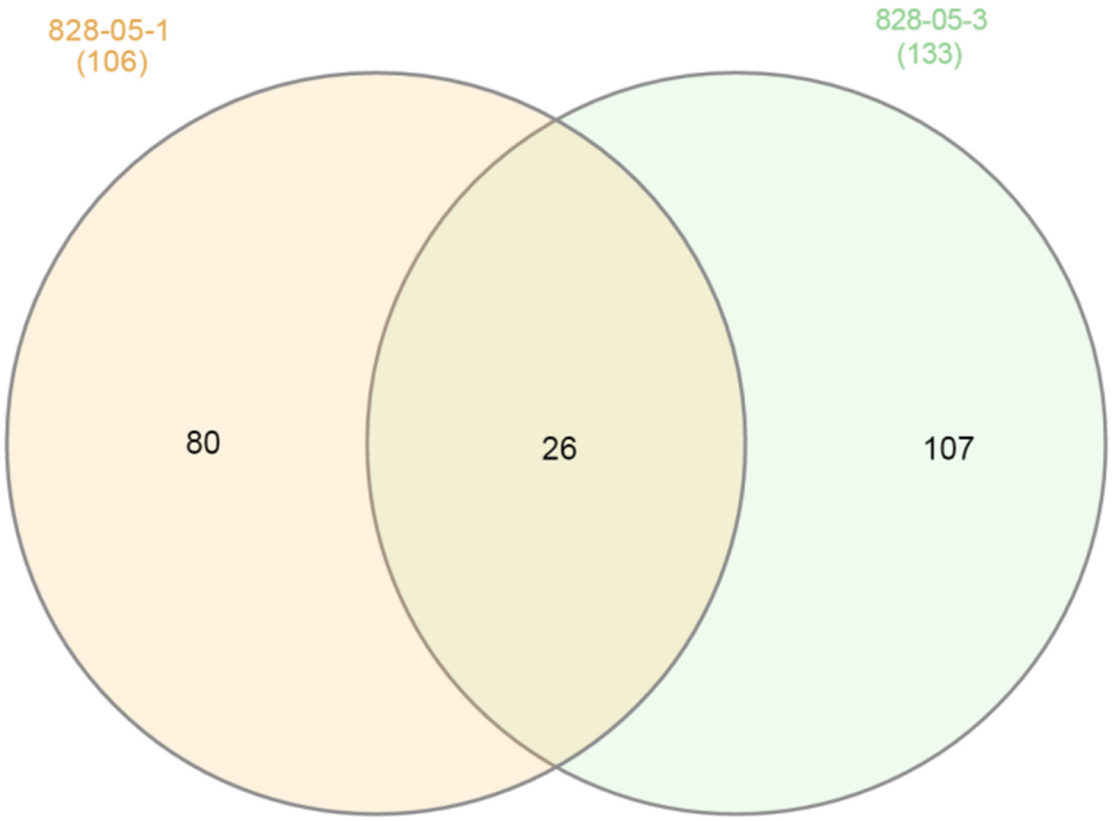


Figure 3.10 Comparison of CNVs inherited by the two half-sibs, exclusively from their sire.

Orange and green circles represent CNVs detected in animals 828-05-1 and 828-05-3, respectively. Numbers in overlapping regions denote the number of CNVs common to both individuals while those in non-overlapping regions are unique for each individual.

Comparison of CNVs between the five sheep revealed that only 75 CNVs (4%) were common to all 5 animals and 57% (1046 out of total 1838) of CNVs were unique to an animal (Figure 3.6). This could be due to huge differences between individuals or low coverage of NGS data which might result in CNV missing during CNV detection. Besides, Figure 3.5 reveals that most of the CNV losses were detected in low depth regions on the chromosomes, which suggests that sequencing depth has a huge influence on CNV detection.

Also, the majority of the CNVR detected in this study were less than 5 kb in size (Figure 3.7). Comparison of the CNVR from this study with those from previous studies in sheep revealed that NGS based CNV detection would provide better resolution (in terms of high CNVR number, but smaller in size) than microarray or aCGH based detections (Table 3.4).

3.5.2 qPCR validation

Leftover DNA (after NGS) was available for only one sheep and the study individuals were no longer alive. Hence, CNV validation, using two pairs of PCR primers, was undertaken on only one animal and the qPCR results corroborated the predicted copy numbers of those two CNVs in the animal. However, such small size of sample for validation reduced the confidence of this study.

3.5.3 Gene annotation

Gene Ontology (GO) and KEGG analysis showed that genes over-represented in the detected CNVRs were associated with brain morphogenesis, the cytoskeleton, cell junctions and calcium ion binding. However, none of those genes passed the threshold for Bonferroni correction for multiple testing which suggests so far the association between these genes and CNVs is unclear.

3.5.4 Comparison with previous Sheep CNV studies

Six papers have been published examining CNVs in sheep (Fontanesi et al. 2011; Hou et al. 2015; Jenkins et al. 2016; Liu et al. 2013; Ma et al. 2015a) and Chapter 2, based on different platforms (aCGH, SNP and NGS) and different genome assemblies (Btau_v4.0, Oar_v1.0, Oar_v3.1, UMD3_OA) (Table 3.4). By comparing the CNVR number and mean and median size, it is clear that CNVnator tended to find smaller CNVR than other algorithms and hence, more CNVR were detected. A detailed analysis was undertaken to overlap the CNVR detected in this study with those from previous studies. The difference in the reference

genome assembly used between studies caused a large problem for the comparisons, especially for CNVR positions, because the positions of SNPs changed across assemblies. Also, the assemblies could not be easily interconverted. Attempts for conversion of CNVR between different assemblies using UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) were unsuccessful, indicating huge differences between assemblies. CNVR results from this study were overlapped with those from Ma et al. (2015a)'s study and our two other studies (Yan et al. 2017 and Chapter 6) that were based on the PennCNV algorithm and employed Oar_v3.1. In order to compare the CNVRs between studies, CNVRs from different studies had to be merged because a single large CNVR detected in a study could be found divided into several smaller CNVRs in another study. This process could cause some bias because some CNVRs were very large and covered several small CNVR from other studies. However, it is a reasonable way to display similarities between different studies.

By comparing the data between NGS, 50K SNP and 600K SNP studies, only four CNVR was common in all studies (Figure 3.11), which indicated that there were huge differences in CNVR discovery based on these three platforms. Furthermore, the average number of CNVR per sample detected using NGS (1,836 CNVRs detected in 5 animals; this study) was much higher than that by 50K SNP microarray (575 CNVRs in 545 animals) (Yan et al. 2017b) or by 600 K SNP microarray (339 CNVRs detected in 93 animals). The ratio of the CNVRs detected in this study (using NGS) overlapping to those detected using 600K SNP microarray (19.2%, 77 out of 399 CNVRs) or 50K SNP microarray (20.7%, 112 out of 539), in our previous studies, was similar. This could indicate that despite being the 600K SNP microarray (606,005 SNPs) more than10 times denser than the 50K SNP microarray (54,241 SNPs) did not improve the quality of CNV calling.

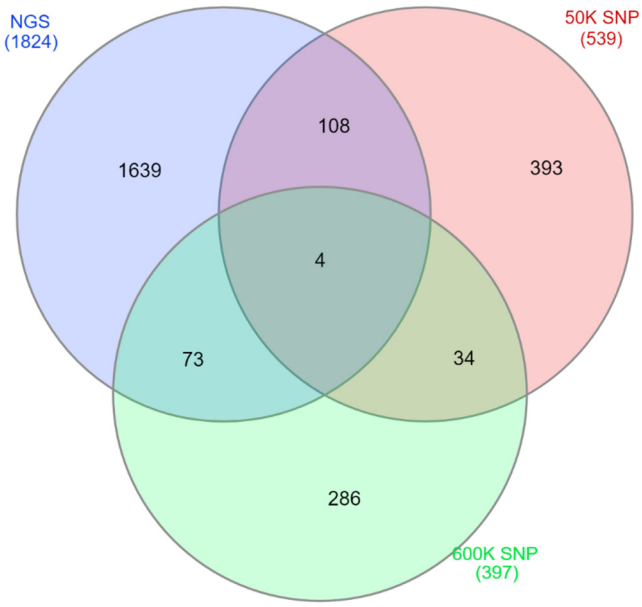


Figure 3.11 Overlap of the CNVRs detected in the current study (based on NGS) with those from previous studies that employed SNP microarrays.

Blue, red and green circles represent CNVRs detected using NGS, 50K SNP microarray and 600K SNP microarray, respectively. Numbers in overlapping regions denote the number of CNVRs common to respective platforms, while those in non-overlapping regions are unique for each platform.

On the other hand, overlapping the CNVRs detected in the current study, using NGS data, with those detected using the 50K SNP data employing three different algorithms (Yan et al. 2017b) revealed that CNVR detected by cnvPartition overlapped better (48.5%, 50 out of 103 CNVRs) than those by PennCNV (28.8%, 120 out of 417) or SVS (18.3%, 130 out of 711). This could be due to the fact that the CNVRs detected by cnvPartition were much longer than those detected by the other two algorithms, so that they had more possibility to overlap with the CNVRs detected using the NGS data. However, comparing the CNVR results from Ma's study (Ma et al. 2015a) and our previous study (Yan et al. 2017b), where both studies employed PennCNV on a 50K SNP microarray platform revealed that only 14 CNVRs of 111 overlapped between the two studies. This indicated that the CNV calling could be influenced

by the populations and breeds sampled. Overall, these comparisons revealed that the detection of CNVR could be influenced by several factors, such as population, breed, platform and algorithms.

Table 3.4 Comparison of the number and size of copy number variant regions (CNVR) detected in this study with those from previous studies in sheep.

Sample Size	CNVR Number	Mean size (kb)	Median (kb)	Size range (kb)	Platform	Algorithms	Assembly	References
11	135	77.6	55.9	24.6-505	Bovine 385k aCGH		Btau_v4.0	(Fontanesi et al. 2011)
329	238	253.57	186.92	13.66-1,300	Ovine 50 K SNP	PennCNV	Oar_v1.0	(Liu et al. 2013)
160	111	123.84	100.53	Unknown	Ovine 50 K SNP	PennCNV	Oar_v3.1	(Ma et al. 2015a)
5	51	304.86		52-2,000	1.4 M aCGH		Oar_v1.0	(Hou et al. 2015)
36	3,488	19		1-3,600	2.1M aCGH		UMD3_OA	(Jenkins et al. 2016)
5	1,836	3.8	1.9	1-73	NGS	CNVnator	Oar_v3.1	This study
385	749	189	118	15.3-6,600	Ovine 50 K SNP	SVS	Oar_v3.1	(Yan et al. 2017b)
	464	305.5	218.1	11.4-2108.8		PennCNV		
	104	1,521.3	395.4	87-12,093.7		cnvPartition		
42	294	22.4	10.1	0.07-198.2	the Ovine Infinium ® HD SNP BeadChip	PennCNV	Oar_v3.1	Chapter 6
51	314	22.4	10.6	0.06-198.2				

3.5.5 Pedigree comparison

In the two offspring studied (828-05-01 and 828-05-03), about 71% and 65.9% of their CNVs, respectively, could be traced from their parents (Figures 3.8 and 3.9), indicating CNVs are strongly inherited. The remainder of the CNVs in the two progenies could be because of random mutation. Also, out of the CNVs inherited by the two half-sibs (106 and

133 CNVs, respectively, by 828-05-01 and 828-05-03), exclusively from the sire, 26 CNVs overlapped (Figure 3.10), further indicating the Mendelian pattern of inheritance of CNV.

3.6 Conclusions

This study successfully detected CNVs in five animals using NGS data. In total 1,836 CNVRs were found in five samples with 1,653 were losses, 181 gains and 2 mixes. The mean and median size was 3,835 and 1,999 bp respectively. There were 587 Ensembl genes located within the identified CNVRs. Compared with previous studies, NGS supports higher resolution so that smaller CNVs were found and has the highest call rate for each individual. Besides, this study also showed that a good reference genome and high sequencing depth are essential for efficient CNV detection using NGS. Finally, a small pedigree comparison gave a clear conclusion that most CNVs could be inherited, but the genetic pattern of some CNVs could be random.

3.7 Acknowledgments

The authors acknowledge provided by the NeSI (www.nesi.org.nz) for high performance computing facilities. In addition, the primary author was supported by Massey University Doctoral Scholarship.

Chapter 4

Genome-wide association study for the associations between CNVs and resistance or resilience to sheep gastrointestinal nematodes

Juncong Yan[1], Hugh T. Blair[1], Andrew Greer[2], Joseph Hamie[2], Patrick Biggs[1] Venkata S.R. Dukkipati[1*]

To be submitted to Journal of Animal Breeding and Genetics

[1] IVABS, Massey University, Palmerston North 4442, New Zealand
[2] Agricultural and Life Sciences, Lincoln University, Lincoln 7647, New Zealand
[*] Correspondence: R.Dukkipati@massey.ac.nz

4.1 Abstract

4.1.1 Background

Gastrointestinal nematodes are one of the most serious parasitic threats for sheep. In order to improve resistance/resilience to nematodes in sheep, genome-wide association study (GWAS) is a good way to identify genetic markers associated with the traits. A GWAS was undertaken, using two kinds of genetic markers, single nucleotide polymorphisms (SNP) and copy number variants (CNV), in Romeny sheep bred for nematode resistance or resilience.

4.1.2 Result

In total, 1,825 CNVs were detected in 53 sheep; these comprised 1,009 losses and 816 gains. The mean and median CNV lengths were 31.5 kb and 15 kb respectively, and the range was from 64 bp to 394.3 kb. One and three CNVRs were found to be associated with live weight and low nematode faecal egg count (FEC) traits, respectively, in CNV-based GWAS. No significant association was found in SNP-based GWAS. In total, seven genes (pertaining to olfactory receptor, neuronal differentiation, putative killer cell immunoglobulin-like receptor and class II histocompatibility antigen) were located within the significant CNVRs. All those CNVRs overlapped with three previously detected quantitative trait loci (QTL). Two of those QTLs were associated with FEC and another with immunoglobulin A level.

4.1.3 Conclusion

A GWAS for three phenotypes (live weight, immunity and FEC) in Romney sheep selected for nematode resistance or resilience revealed one and three CNVRs to be significant for phenotypes, immunity and FEC, respectively. Seven genes and three previously reported QTLs were found to be located within the significant CNVRs. No SNPs were found significant at genome- wide scale, probably due to very small sample size.

Since there was no overlap between CNV and SNP based GWAS results, SNPs and CNVs could represent different aspects of genetics involving quantitative traits.

4.1.4 Keywords: sheep, GWAS, SNP, CNV, nematodes

4.2 Introduction

Gastrointestinal nematodes are one of the most serious parasitic threats for sheep (Familton and McAnulty 1997; Perry and Randolph 1999), costing approximately $300 million annually to the New Zealand sheep industry (Rattray 2003). Regular drenching of sheep with anthelmintics could lead to development of anthelmintic resistance. In addition, there is increased consumer preference for chemical-free products, such as meat. Therefore, alternative anti-parasite strategies are necessary. Genetic selection is an important animal husbandry tool to improve the quality of domestic animals. Several studies have shown that resistance to nematodiasis in sheep is highly variable and heritable so that selective breeding is an alternative choice for nematode control (Morris et al. 1995; Morris et al. 2000; Morris et al. 2005).

Phenotypes, such as resistance/susceptibility to parasites, have been found to be associated with several genes rather than just one (Marshall et al. 2009). Therefore, the old, single gene basded research method does not work anymore. In this situation, genome-wide association study (GWAS) provides a path to deal with this issue. Simply, GWAS is a kind of statistical tool to find association between genetic markers across the genome and the trait in question. The first successful GWAS was undertaken by Klein et al. (2005) in human, who found a single nucleotide polymorphism (SNP) variant in the complement factor H gene (CFH) to be strongly associated with age-related macular degeneration (AMD) in humans. Currently, GWAS is being widely used in animal breeding. For instance, quantitative trait loci (QTL) for carcass weight in cattle (Nishimura et al. 2012) and growth and meat production traits of sheep (Zhang et al. 2013) have been identified by GWAS.

GWAS can be undertaken based on several kinds of genetic marker data, with SNP data being the most popular. However, SNPs do not explain the all genetic variation, such as structural variation. (Manolio et al. 2009).

Recent developments in genome research have identified new types of genetic variation such as Copy Number Variation (CNV), where large segments of DNA are either duplicated or deleted. Associations of CNVs with complex human diseases have been reported. In 2004, Iafrate et al. and Sebat et al. found large-scale CNV in humans and thought that those genetic variations could reveal associations between genotype and phenotypes such as susceptibility to disease and regulation of cell growth (Iafrate et al. 2004; Sebat et al. 2004). Two years later, Redon et al. published the first draft of CNV in human, which had 1447 CNV regions (CNVRs), covering 12% of the genome (Redon et al. 2006). A study in humans revealed that CNVs could capture up to 17.7% of the genetic variation for gene expression, having no overlap with that explained by SNPs (Stranger et al. 2007). Hence, CNV would be a good choice in GWAS to reveal associations between genotype and phenotype. However, so far there is no CNV-based GWAS reported in sheep.

The objective of this study was to investigate the use of SNP and CNV as genetic markers in a GWAS analysis to identify the association between genotypes and either resistance or susceptibility to gastrointestinal nematodiasis in sheep.

4.3 Materials and methods

4.3.1 Ethics statement

This study was carried out following guidelines of the 1999 New Zealand Animal Welfare Act and was approved by Lincoln University Animal Ethics Committee (Permit Numbers: LUAEC#588).

4.3.2 Tissue sampling, genotyping and phenotypes

Ear punch samples were collected into an Allflex tissue sampling unit (TSU), using an Allflex® NZ tissue sampling applicator (TSU Applicator – 22134). Samples were obtained from 53 Romney sheep belonging to two selection lines; nematode resistant, n = 21, and nematode resilient, n = 32, currently being maintained at Lincoln University, New Zealand. These two lines were selectively bred for over 16 generations (1985-2009) for resistance and resilience to gasto-intestinal nematodes (GIN), based on faecal egg count (FEC) using best linear unbiased prediction (BLUP) techniques. Details regarding the selection lines were described elsewhere (Baker et al. 1990). The original "susceptible" line in (Baker et al. 1990) and other papers published by Agresearch, has been referred to as "resilient" line in this thesis, based on the ability of the animals in this line to be resilient in terms of production, despite exhibiting FEC.

The tissue samples were submitted to AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand, for DNA extraction and SNP genotyping using the Ovine Infinium® HD SNP BeadChip (Additional file: Table S4.1 Sample information).

Twenty-six rams and 27 ewe lambs were grazed separately in existing paddocks containing predominantly ryegrass pasture to allow for natural infection with infective stage larvae (L3) parasite of mixed species. From weaning at a mean 92 days-of-age, animals were sampled approximately every 10 days until 341 days-of-age. Faecal samples were collected from the rectum of each lamb immediately upon yarding for the determination of the concentration of nematode eggs in the faeces using a modification of the McMaster method by floatation in saturated sodium chloride solution as described by Whitelock (1948) with a sensitivity of 100 eggs per gram (EPG). Saliva samples were taken using mouth swabs that were then centrifuged at 1200 g and the saliva stored at -20^{o}C until analysis. Animals were then fasted

without access to feed or water for 16 h before the recording of fasted live weights after which they were returned to grazing. Live weights (LW) were recorded with the use of electronic identification tags (Allflex New Zealand) and an Aleis tag reader connected to a semi-automated Prattley autodrafter with a sensitivity of 0.2 kg.

Saliva samples were analysed for antibody to GIN larvae using an enzyme linked immunosorbant assay (ELISA) similar to that described by Douch et al., (1994). Fifty µl of Trichostrongylus colubriformis L3 antigen/well at 2 µg/ml in coating buffer (stock=300 µg/ml = >1:150 dilution) were sealed and incubated at 4ºC overnight. The ELISA plates were washed 5 times with dilution buffer containing 0.1% (w/v) Tween 20 (W-T20). Then, 200 µl/well blocking buffer (10mM-phosphate buffer at pH 7.2 containing 0.5% Tween-20 and 5% bovine skim milk powder) was added to plates and incubated for 2 h at room temperature. Plates were washed 5 times with washing buffer solution. Diluted saliva (1:10 for IgA and 1:100 for IgG) was added to ELISA plates at 50 µl /well, incubated for 2 h and then plates were washed 5 times with washing buffer solution. Rabbit anti-sheep IgG and or IgA conjugated with horseradish peroxidase (Bethly Laboratories inc., USA), diluted 1:2000 with ELISA buffer, was added to each well (100µl) and incubated for 1 h at room temperature. Plates were washed 5 times with washing buffer solution. To develop colour, 100 µl/well of tetramethyl benzidine (TMB) substrate was added and incubated for 40 minutes at room temperature. The substrate 0.05M phosphate-citrate buffer pH 5.0 was made of 25.7ml 0.2M Na_2HPO_4 + 24.3ml 0.1M citrate and made up to 100ml with deionised distilled water (dH_2O) to which 2 µl of 30% H_2O_2 + 1 TMB tablet per 10ml buffer were added. The reaction was stopped by adding 100 µl/well of stop solution (6.9ml of 1.25M concentrated H_2SO_4) and then plates read for optical density at 450nm using ELISA plate reader.

Based on the phenotypes recorded, the 53 sheep were classified in three different ways, based on live weight, immunity and FEC: a) resilient (cases) vs non-resilient (controls), b) resistant (cases) vs non-resistant (controls) and Low FEC (case) vs High FEC (controls). For resilience, lambs were identified as to whether they were above or below average for their sex for mean live weight (LW), live weight gain and cumulative growth. If they were above average for any two of these criteria they were considered resilient. For the ewes (adult LW) they were considered resilient if their LW at tailing was above average. For the immunity, animals were considered immune (resistant) if their mean FEC was lower, and IgG and IgA levels were greater than the average for their sex. Again, if they had two out of the three they were considered to be immune (resistant). In case of FEC, the animals with <100 EPG were considered low FEC, while those with greater than 100 EPG as high FEC.

4.3.3 Quality control and CNV detection

The Ovine Infinium® HD SNP BeadChip was designed based on Oar_v3.1 gene map. The original SNP data (idat files) was converted to ped and map file from GenomeStudio® using PLINK Input Report Plug-in v2.1.4. Then, quality control of SNP was done based on call rate of samples, Minor Allele Frequency (MAF) and Call Frequency (Call Freq) using PLINK1.9 beta. SNPs with call rate less than 99% or MAF less than 1% were excluded. Individual sample was excluded if the overall SNP call rate was less than 97%. Additionally, SNPs that were not in Hardy-Weinberg equilibrium (HWE; $p<10^{-6}$) were also excluded.

CNVs were detected using PennCNV v1.03 at 3 minSNPs (Wang et al. 2007). PennCNV employs an integrated hidden Markov model on an Illumina platform. Based on three criteria, population frequency of B allele (PFB), SNP genome coordinates and a trained hidden Markov model (HMM) file, the most likely state-transition path could be analysed using the Viterbi algorithm. A PFB file of SNPs was created using the compile_pfb.pl program in PennCNV, based on SNP data from 93 New Zealand Romney (chapter 6). The GCmodel

option of PennCNV was not applied in this study because this model was not yet optimised for non-human species (Wang K, personal communication). Signal intensity files, which had Log R ratio (LRR) and B Allele Frequency (BAF), were created from the final report from GenomeStudio® using a PennCNV plugin, split_illumina_report.pl. PennCNV integrates LRR, BAF and PFB for each SNP, and the distance between adjacent SNPs, into a HMM, for detecting CNV. Final quality control of the detected CNVs was done using a program, filter_cnv.pl , of PennCNV software (Wang et al. 2007). Identity by descent (IBD) test was done to evaluate the genetic distance between each two samples using plink1.7.

4.3.4 qPCR validation

Four selected CNVs in eight individuals were validated by qPCR, using StepOnePlus™ Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) (Ma and Chung 2014). The *DGAT1* gene was used as reference since it was shown to be free from copy number variation (Fontanesi et al. 2011). Details of primer sequences, target regions in the sheep map, as well as PCR conditions are shown in Additional files: Table S4.2 qPCRresult.

The copy number of the amplified regions was calculated by a relative standard curve method (Biosystems 2004) as follow:

$Qty = 10^{\frac{Ct-b}{m}}$, where *Qty*, *Ct*, m and b are the relative quantity of amplified fragment, threshold cycle, slope and y-intercept of the standard curve.

$$copy\ \ number = \frac{Qty(\text{NormalizedTarget})}{Qty(\text{NormalizedReference})} = \frac{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{target sample}}{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{reference sample}}$$

However, it is difficult to find a standard sample as a reference which has copy number variation. Therefore, firstly, the reference was selected randomly. The copy number of reference was assumed as 1 copy, then calculate the accuracy of qPCR. After that, the copy

number of reference was assumed as 2, and 3 and did the same process again. Because the gene has more than 4 copies is rare, no more assumption was set up. By comparing the accuracy between the copy numbers 1, 2, 3 the copy number which has highest correction rate is considered as the correct copy number. Of course, this method could have bias since the assumption of copy number of reference could be wrong. The copy number evaluation thresholds table is given below (Table 4.1). Based on hypothetical copy number, using the copy number value calculated by above equation of each sample, the actual copy number of each sample can be evaluated.

Table 4.1 Hypothetical copy numbers of the reference and their thresholds (based on qPCR) for copy number evaluation

Hypothetical copy number of the reference sample	**1 copy**	**2 copies**	**3 copies**	**4 copies**
1 copy	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5
2 copies	0.25-0.75	0.75-1.25	1.25-1.75	1.75-2.25
3 copies	0-0.459	0.459-0.825	0.825-1.165	1.165-1.495

4.3.5 Genome-wide association study (GWAS)

GWAS, based on SNPs as well as CNV, was undertaken on the three phenotypes, live weight (22 cases /31 controls), resistance (24cases/29 controls) and FEC (31 cases/22 controls). GWAS based on SNPs was done using PLINK1.9 beta (Purcell et al. 2007) based on a 1 df chi-square allelic test (Equation 4.1).

Equation 4.1 Basic Allelic Test for association by chi-square allelic test (Clarke et al. 2011).

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(m_{ij} - E[m_{ij}])^2}{E[m_{ij}]}$$

$$E[m_{ij}] = \frac{m_{i.}m_{.j}}{2n}$$

Each of the three genotypes at a SNP locus, aa, aA, and AA, contain different numbers of the two alleles, a and A, which formed a 2X2 table. m_{ij} represents the number of SNP allele (a or A) in two groups (cases or controls) x^2 represents null hypothesis of no association under 1 d.f. (Degree of freedom).

Allele	**a**	**A**	**Total**
Cases	m11	m12	m1.
Controls	m21	m22	m2.
Total	m.1	m.2	2n

PLINK 1.07 was used to do GWAS based on CNV data. The basic idea was same as above, but CNVs were treated as 0, 1, 2, 3, based on their event.

Three binary phenotypes, resilience, resistance and FEC, were applied. The command: plink1.9 --file 600KQC --pca --keep 53samplelist.txt was applied to calculate principal components analysis (PCA) which was used to correct effect of population stratification. After calculation, the SNP markers with a p value less than $1X10^{-5}$ and $5X10^{-8}$ were considered significant at individual SNP scale and genome-wide scale, respectively (Turner 2014; Fadista et al. 2106). The CNVs with an EMP1 (Empirical p-value for individual CNV test) less than 0.05 were considered to be significate at individual CNV level, while EMP2 (Empirical p-value, corrected for all tests) less than 0.05 was considered significant at genome-wide scale. Significant CNVs that overlapped by at least one bp were merged as a significant CNVR. The quantile-quantile (Q-Q) plot was done using R package, snpStats (Clayton 2012) to evaluate the accuracy of GWAS and Manhattan plot was made by R package, qqman, to show the in genome wide SNPs distribution (Turner 2014).

4.3.6 Gene annotation

Using a custom written script in Perl and SQL (Appendix 4.1), position information of CNVs and SNPs, that were significant, were matched to assembly Oar_v3.1 to get the names of genes (Ensemble database). Then all gene names were input into online tools, bioDBnet (biodbnet-abcc.ncifcrf.gov/db/db2db.php), to get information of gene function.

4.4 Results

4.4.1 Quality control

Only 477,531 out of the total 606,005 SNPs located on the 26 autosomes passed the quality control threshold and were retained for further analysis. All 53 sheep (21 resistance and 32 resilience) passed the quality control for PennCNV (Additional file: Table S4.1).

The IBD test showed 485 sample pairs of total 1378 (35.1%) are too close (IBD >0.2). Based on limited sample size, it would not be used to exclude sample for further analysis.

4.4.2 CNV detection

In total, 1, 825 CNVs were detected in 53 sheep, these comprised 1, 009 losses and 816 gains. The mean and median CNV lengths were 31.5 kb and 15 kb respectively, and the range was from 64 bp to 394.3 kb (Additional file S4.1).

Table 4.2 Results of qPCR validation of four randomly selected CNVs.

Primer ID	Samples validated	Validation accuracy	
		Proportion	Percentage
J26	6	6	100.0%
J27	8	4	50.0%
J28	8	5	62.5%
J29	8	7	87.5%
Total	30	24	83.3%

4.4.3 qPCR validation

Four randomly selected CNV segments were validated by qPCR, using DNA from eight samples. Validation accuracies for the four markers ranged between 50% to 100%, with an overall mean of 75% (

Table **4.2**; Additional file: S4.2 qPCR result).

4.4.4 Genome-wide association

Obvious population stratification was found (Figure 4.1) between resilience and resistance. Elbow point (PC4), the first non-significnat component, was decided using an eigenvalue plot (Figure 4.2).

Q-Q plots were made to show the accuracy of each phenotype before and after PCA correction. The plot of live weight (Figure 4.3) shows that the realistic distribution of SNPs matched the expected distribution and the influence of PCA correction was small, therefore the PCA was not be applied for further analysis of this phenotype. On the other hand, the plots of immunity and FEC (Figure 4.4, Figure 4.5) reveal less disparity between the realistic and expected distributions of SNPs post-PCA correction, compared to that seen prior to PCA correction. Therefore, PCA correction was applied for further analysis.

Only three and one CNVRs were found to be significantly (EMP2 < 0.05), associated with the phenotypes, live weight and FEC based on CNV data (Table 4.3). At individual CNV level (EMP1<0.05), about 4, 12, 17 CNVRs were significant for live weight, FEC and resistance traits, respectively. No SNPs were significant at genome wide scale ($p < 5 * 10^{-8}$) in this study (Figure 4.6 – 4.8) (Additional file: Table S4), but two SNPs, oar3_OAR3_21838826, oar3_OARX_116590638 were significant at single SNP scale in case of live weight.

4.4.5 Gene annotation

Based on significant SNPs, one and three significant CNVRs were found in live weight and FEC (Table 4.3). One of them are shared between two GWAS analysis which is located on chromosome 3, from 164580644 to 164730778, coding 4 genes about olfactory receptor 9K2-like and neuronal differentiation 4. This CNV also is overlapped on a strong FEC associated QTL zone, QTL:12891.

Other two CNVRs just were found in CNV based GWAS analysis for FEC. One is on chromosome 14 from 59,866,958 to 59,902,608, coding one gene about putative killer cell immunoglobulin-like receptor like protein KIR3DP1 and is also overlapped on a nematodirus FEC associated QTL zone, QTL: 12,893 while the other is on chromosome 20 from 25,402,869 to 25,449,539, coding two genes about SLA and boLa class II histocompatibility antigen. It is not located on any parasite associated QTL zone directly, but located on an immunoglobulin A level associated QTL zone, QTL 12896 (Additional file: Table S4).

Table 4.3 Significant (EMP2<0.05) CNVRs detected by GWAS for live weight and FEC and gene annotation

Live weight			
Significant CNVR	**Size**	**Gene**	**Annotation**
3:164,580,644-164,730,778	150.1kb	LOC101118264	olfactory receptor 9K2-like
		LOC101120067	olfactory receptor 9K2-like
		LOC101118521	olfactory receptor 9K2-like
		NEUROD4	neuronal differentiation 4
FEC			
Significant CNVR	**Size**	**Gene**	**Annotation**
3:164,580,644-164,730,778	150.1kb	LOC101118264	olfactory receptor 9K2-like
		LOC101120067	olfactory receptor 9K2-like
		LOC101118521	olfactory receptor 9K2-like
		NEUROD4	neuronal differentiation 4
20:25,402,869-25,449,539	46.6kb	LOC101109220	SLA class II histocompatibility antigen
		LOC105603927	boLa class II histocompatibility antigen
14:59,866,958-59,902,608	35.6kb	LOC101104523	putative killer cell immunoglobulin-like receptor like protein KIR3DP1

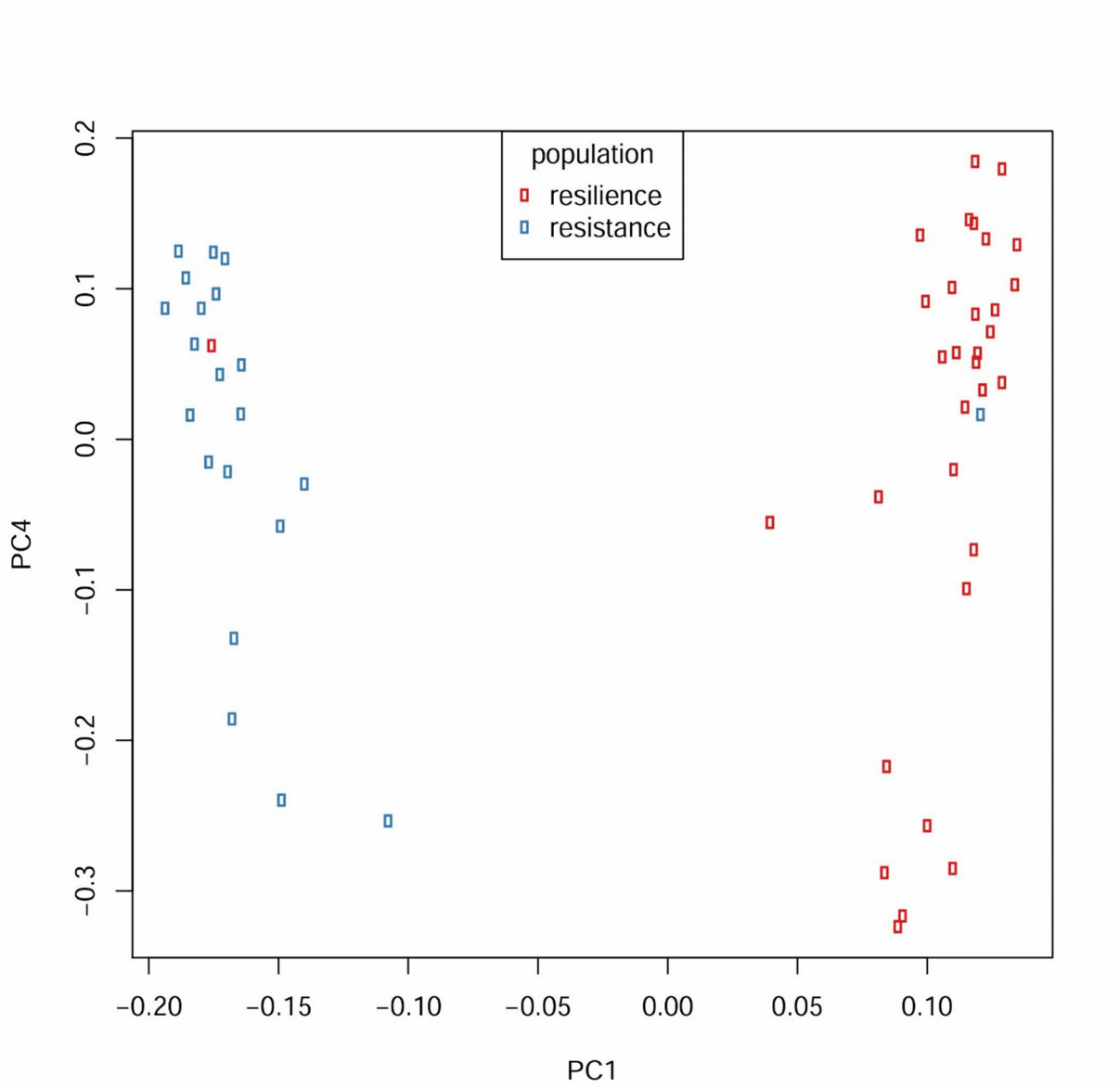


Figure 4.1 Principal component analysis revealing population stratification.

The red and blue box represents resilience and resistance. The X and Y axis represent principal component 1 (PC1) and principal component 4 (PC4).

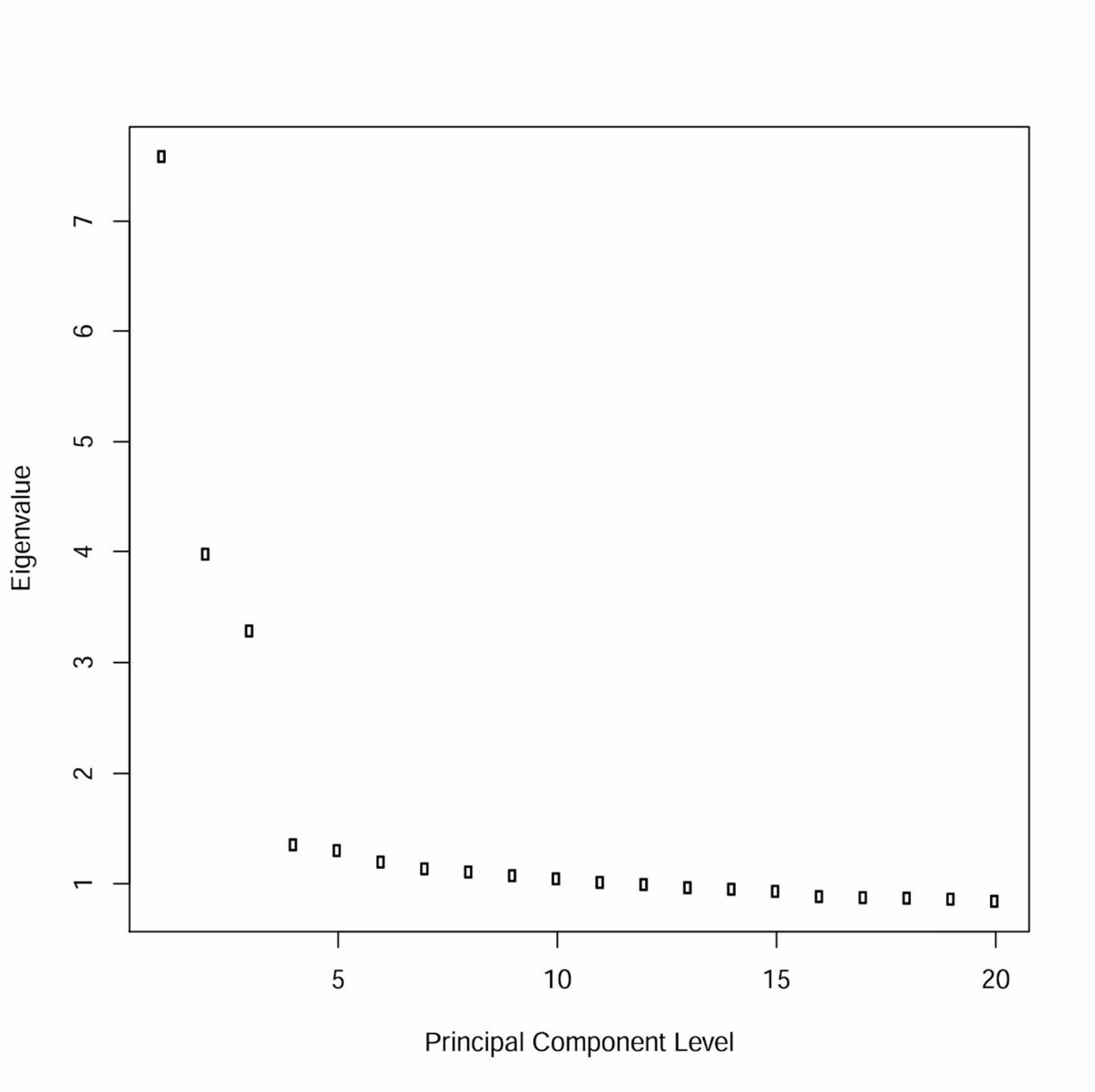


Figure 4.2 Principal component analysis - eigenvalue plot

The X and Y axes represent principal component level and eigenvalue, respectively. The eigenvalues for principal components 1 to 6 are 7.58258, 3.9762, 3.28293, 1.35045, 1.29684, 1.19197, respectively.

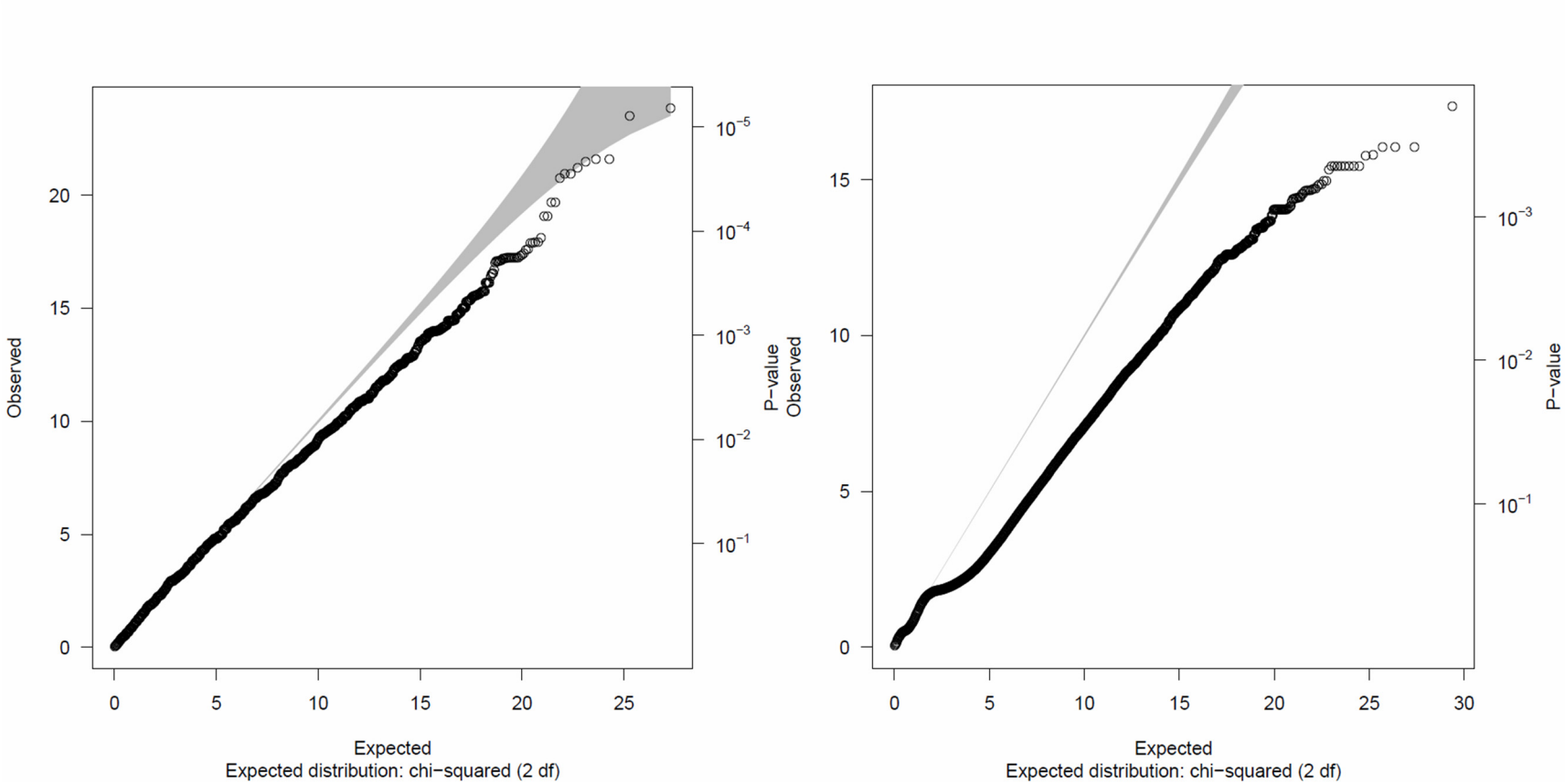


Figure 4.3 Q-Q plot for before PCA correction (left) and after PCA correction (right) of SNP's GWAS for live weight.

Circular points represent realistic distribution of SNPs and shadow represents expected distribution.

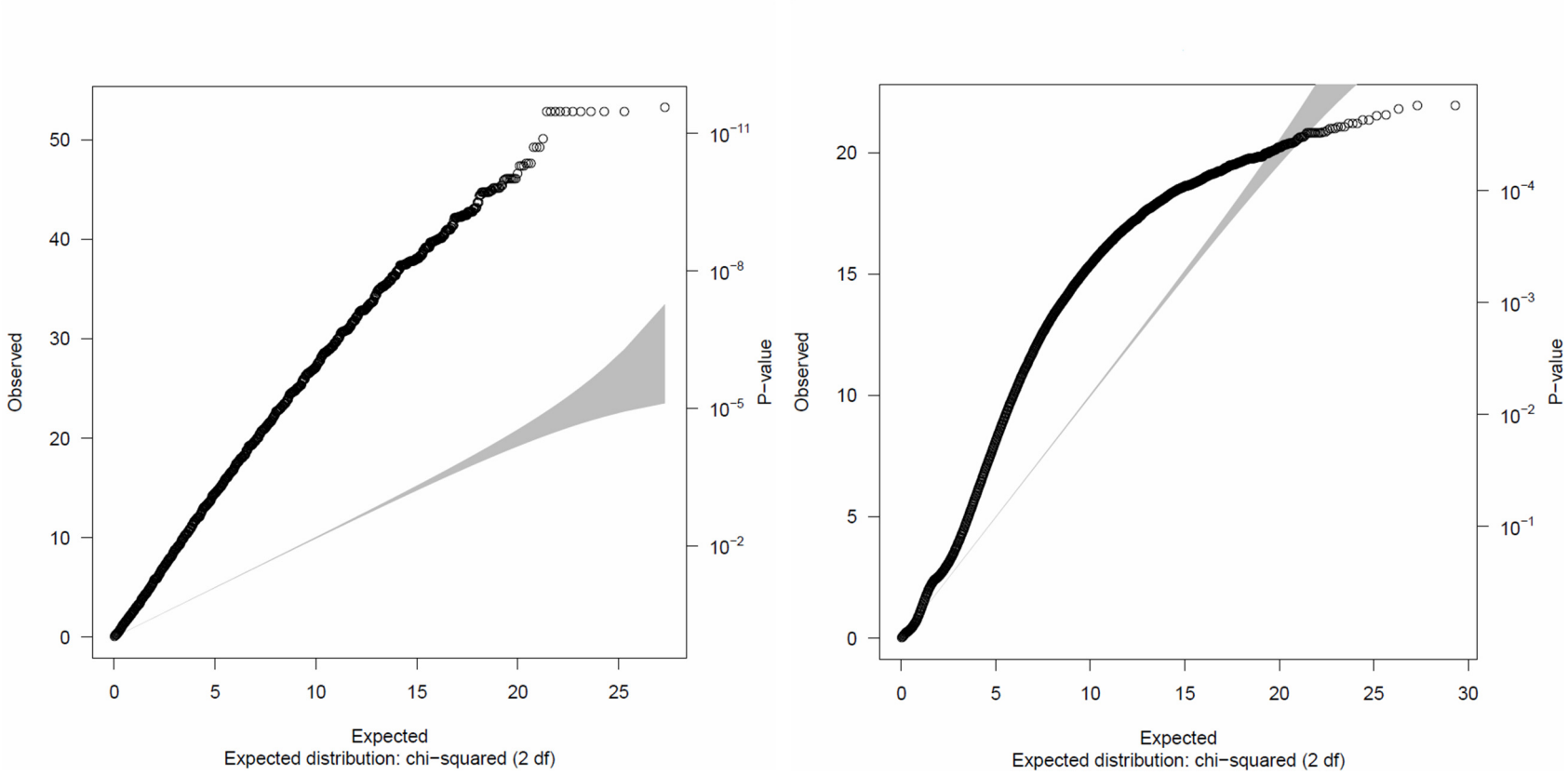


Figure 4.4 Q-Q plot for before PCA correction (left) and after PCA correction (right) of SNP's GWAS for immunity.

Point represents realistic distribution of SNPs and shadow represents expected distribution.

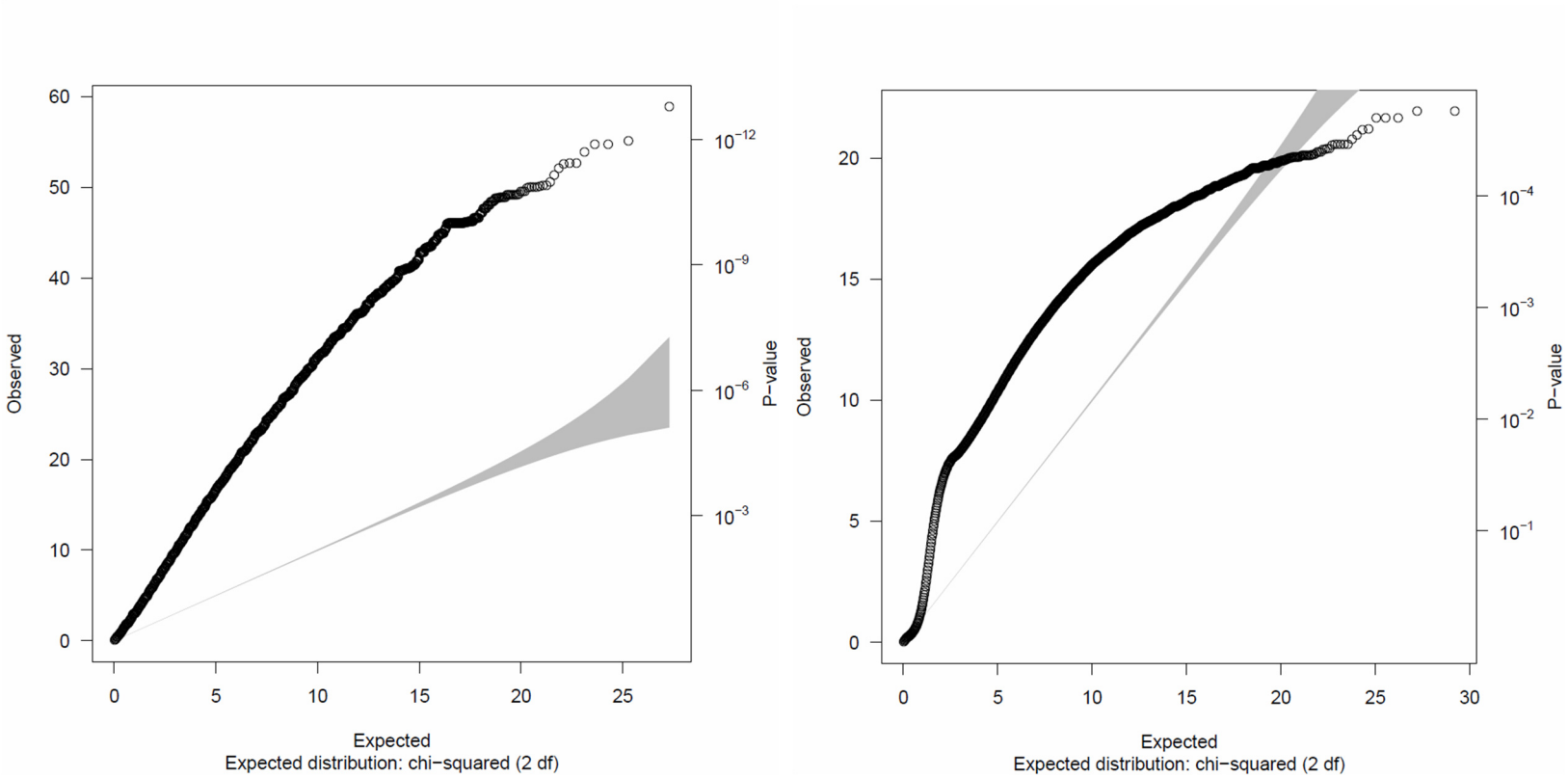


Figure 4.5 Q-Q plot for before PCA correction (left) and after PCA correction (right) of SNP's GWAS for FEC.

Point represents realistic distribution of SNPs and shadow represents expected distribution.

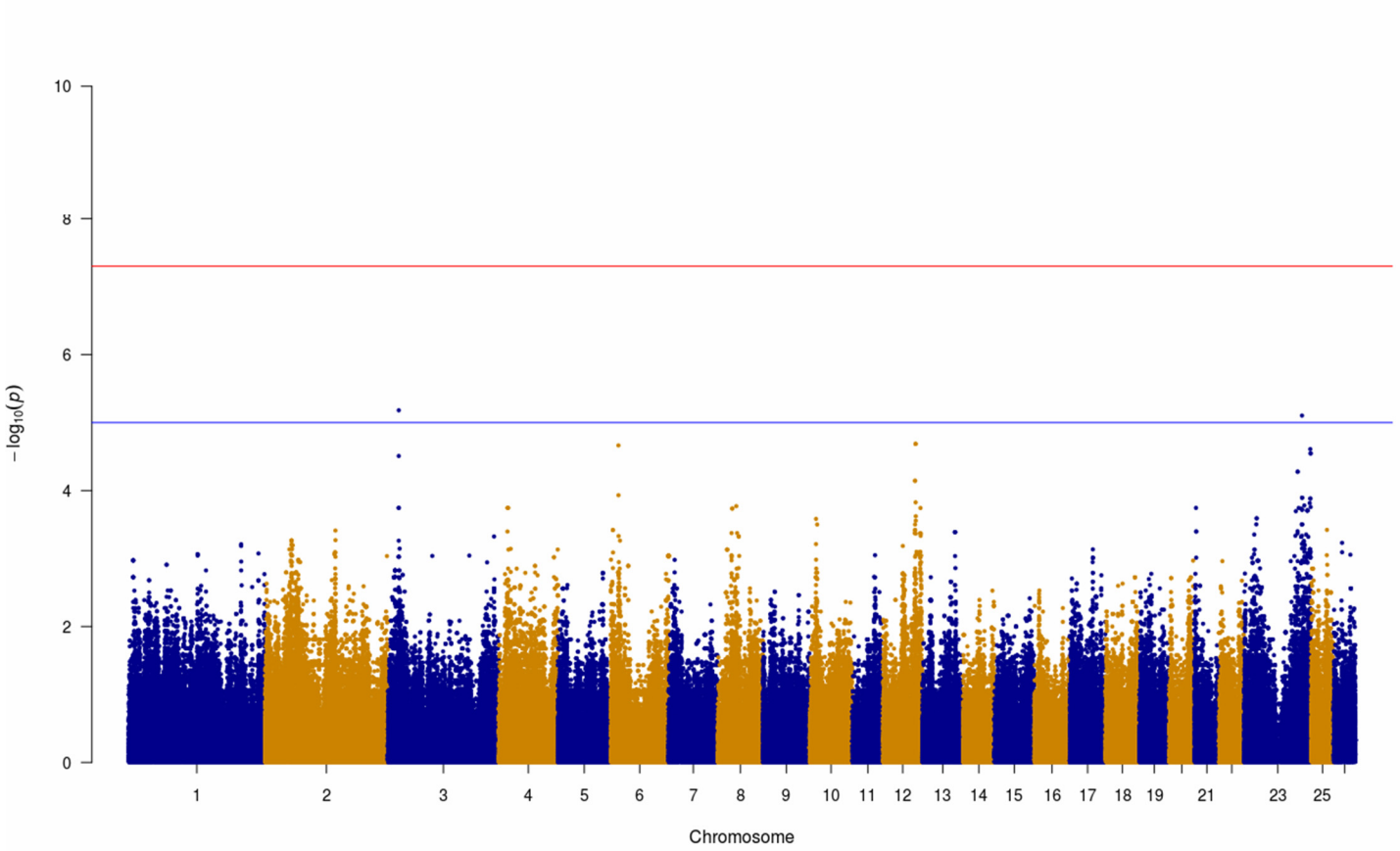


Figure 4.6 Basic Allelic Test for association by chi-square allelic test for live weight.

Points represent SNPs. X and Y axes represent chromosomes and –log10 (*p*), respectively. The blue and red lines represent the thresholds of significance at individual and genome-wide scale.

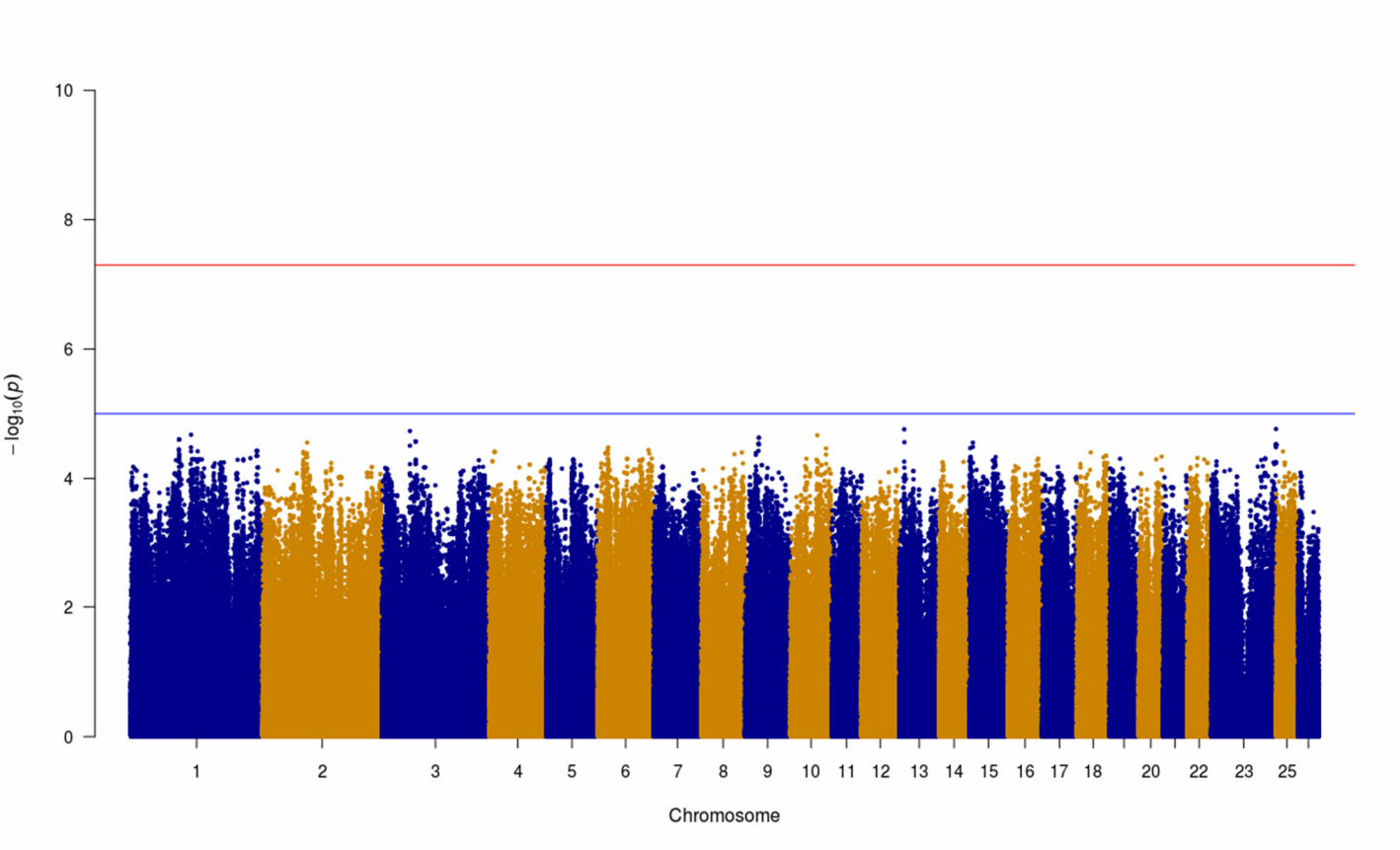


Figure 4.7 Basic Allelic Test for association by chi-square allelic test for immunity.

Points represent SNPs. axes represent chromosomes and –log10 (p), respectively. The blue and red lines represent the thresholds of significance at individual and genome-wide scale.

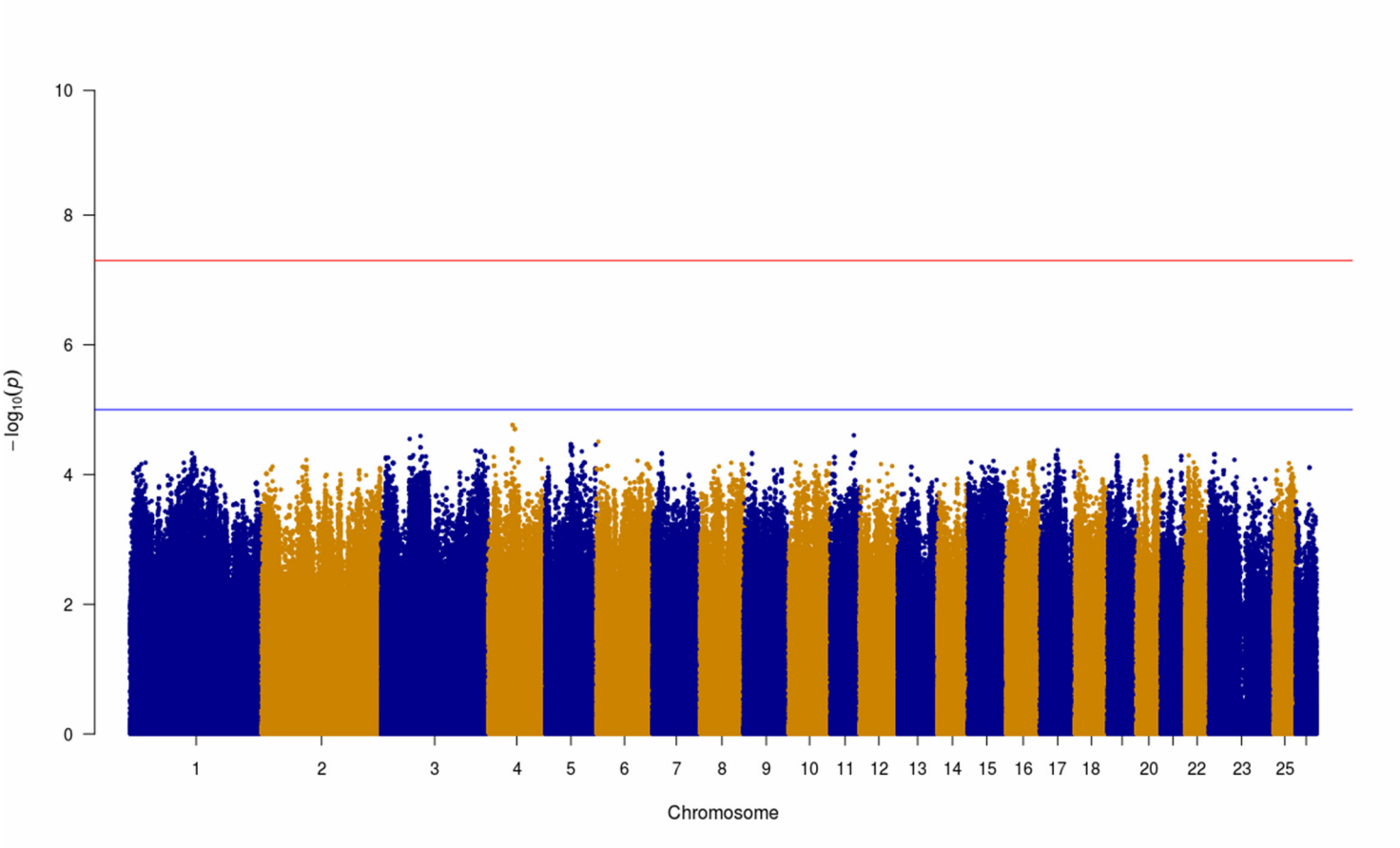


Figure 4.8 Basic Allelic Test for association by chi-square allelic test for FEC.

Points represent SNPs. axes represent chromosomes and –log10 (*p*), respectively. The blue and red lines represent the thresholds of significance at individual and genome-wide scale.

4.5 Discussion

In this study, CNV based GWAS found one and three significant (EMP2 < 0.05) CNVRs for live weight and FEC, respectively, while no SNPs were found significant.

There was no overlap between the results of GWAS based on CNV and on SNPs which indicated that CNV and SNP reflected different aspects of genetic diversity.

One of CNVRs, chromosome 3:164,580,644 -164,730,778, is shared between live weight and FEC phenotype, involving four genes, about olfactory receptor 9K2-like (LOC101118264, LOC101120067, LOC101118521) and neuronal differentiation 4 (NEUROD4). Olfactory receptor gene was also found in a previous CNV in cattle (Hou et al. 2012c) to be associated with gastrointestinal resilience. Besides, a strong FEC associated QTL, QTL: 12891(Davies et al. 2006), was found in this CNVR.

Other two CNVRs were found in CNV based GWAS analysis for FEC. One is on chromosome 14 from 59,866,958 to 59,902,608, coding one gene (LOC101104523) about putative killer cell immunoglobulin-like receptor, which is associated to a kind of parasite, malaria (Yindom et al. 2012). Besides, it is also overlapped on a nematodirus FEC associated QTL zone, QTL: 12893 (Davies et al. 2006).

The third CNVR is on chromosome 20 from 25,402,869 to 25,449,539, coding two genes about SLA (LOC101109220) and Bola (LOC105603927) class II histocompatibility antigen. It is not located on any parasite associated QTL zone directly, but located on an immunoglobulin A level associated QTL zone, QTL 12896 (Davies et al. 2006) which has been found to be associated to resistance or resilience of intestinal nematodes in ruminant (Paterson 1998).

In the current study, no SNPs were significant at genome-wide scale ($P<5X10^{-8}$), which could be due to small size of population. A previous study (Guðmundsdóttir et al. 2015) has proved

that microarray based case-control GWAS required at least 903 samples in each group in order to generate statistically significant results of association with a study power of 80%. Only two SNPs passed individual SNP-scale threshold. Those two SNPs were found to be located in no-gene zone for live weight, but one of them, oar3_OAR3_21838826, was found located in FEC associated QTL, QTL:14155 (lower egg counts for *Trichostrongylus colubriformis*) (McNally and Murrell 2010). Besides, IBD test showed that the affinity relationship of these samples to be quite close (the IBD of 35.1% sample pairs > 0.2). This might have influenced the result of GWAS.

4.6 Conclusion

This study looked at genome-wide association of CNV and SNPs for three different phenotypes (live weight, immunity and FEC) in Romney sheep. Only one and three CNVRs were identified to be significantly associated with live weight and FEC, respectively. In total, seven genes were located in these CNVs, involving olfactory receptor, neuronal differentiation, putative killer cell immunoglobulin-like receptor and class II histocompatibility antigen. Besides, all these CNVRs overlapped with three previously reported QTL zones. On the other hand, no SNPs were found significant at genome-wide scale, probably due to very small sample size and closed genetic distance.

There was no overlap between CNV and SNP based GWAS results, showing that SNP and CNV could represent different aspects of genetics.

4.7 Additional files

Additional file: Table S4.1

Chapter 5

Somatic mosaicism of copy number variation in sheep using Ovine Infinium® HD SNP BeadChip

Juncong Yan[1], Hugh T. Blair[1], Patrick J. Biggs[1], Sarah J. Pain[1], Venkata S.R. Dukkipati[1*]

To be submitted to PLOS ONE

[1] IVABS, Massey University, Palmerston North 4442, New Zealand
[*] Correspondence: R.Dukkipati@massey.ac.nz

5.1 Abstract

5.1.1 Background

Somatic mosaicism (SM) is existence of different genotypes at a given locus in two or more populations of cells in an individual. These kinds of mutation have been associated with many kinds of diseases in human. Recent advances in genome-wide single nucleotide polymorphism (SNP) microarray technology give opportunities to expand this knowledge from humans to animals and to provide evidence to understand structural and functional differences between tissues. In this study, SM was probed in tissues sampled from adult and foetal sheep, using two copy number variants (CNV) detection algorithms (cnvPartition and PennCNV) on an Illumina high density SNP microarray platform.

5.1.2 Results

In total, 1546 and 1764 CNVs were detected using cnvPartition and PennCNV respectively. Of these, 693 (44.8%) and 944 (53.5%) CNVs were found to exhibit mosaicism. Overlapping the results of the two algorithms revealed that only 664 CNVs (441 losses and 223 gains) were detected by both algorithms, which indicated CNV detection differed significantly between the algorithms. Only 174 CNVs (26.2%), detected by both algorithms exhibited somatic mosaicism. Comparison of CNVR detected between adults and foetuses revealed that only 45 out of total 140 (32.1 %) CNVRs were common while only six CNVRs were found in all tissues, while 55 were tissue-specific indicated that age and tissue differentiation could be a factor influencing the formation of CNV.

5.1.3 Conclusion

This study showed that significant mosaicism of CNV exists in sheep and mosaicism was influenced by age, individuals, CNV detection algorithm and tissue analysed. Employing a

combination of CNV detection algorithms, rather than individual algorithms, is crucial in order to achieve a sufficiently high accuracy to estimate somatic mosaicism.

5.1.4 Keywords: Somatic mosaicism, CNV, SNP, sheep

5.2 Introduction

Somatic mosaicism (SM) are mutations occurring in specific cells of an individual so that two or more populations of cells with different genotypes are found (Biesecker and Spinner 2013; Edwards 1989; Freed et al. 2014; Strachan and Read 2011). Mosaicism can happen in both somatic and germline cells, those occurring in the latter can be inherited by the next generation. The mutations that occur in offspring, but not found in parents are called *de novo* mutations (Poduri et al. 2013).

Detection of SM can be traced back to 1914, when abnormal karyotypes were detected in cancer tissue by Theodor Boveri (1929). After that, only a few studies were undertaken in this area until the 1970s and 1980s, when the relationship between somatic cell gene rearrangement and the functional diversity of immunoglobulin was discovered (Brack et al. 1978; Tonegawa 1983). During the past decade, advances in molecular techniques have enabled the discovery of an association of SM with cancer (Vogelstein et al. 2013), neurological disease (Poduri et al. 2013), autism (Sanders et al. 2012) or ageing (Hoeijmakers 2009; Kennedy et al. 2012).

SM has been reported in plants (Gill et al. 1995) and animals (Schaible 1963). To date, the most studies have examined humans. Therefore, there is an opportunity to expand knowledge of somatic mosaicism from humans to animals.

There are three methods for the detection of genome-wide SM: array comparative genome hybridization (aCGH), single nucleotide polymorphism microarray (SNP microarray) and

next-generation sequencing (NGS). aCGH uses competitive hybridization of fragmented target DNA and control DNA marked with different fluorophores to detect CNV while SNP microarray uses the intensity signals of a number of SNP probes to predict CNV. Next generation sequencing (NGS) (Reis-Filho 2009) is based on shotgun sequencing technology (Fleischmann et al. 1995). Compared to first generation sequencing based on Sanger's method, NGS can produce a massive amount of genetic information in a short time.

aCGH was the first microarray based technology used for detecting SM, it can identify mosaicism when the variant cell proportion is greater than 10 percent (Biesecker and Spinner 2013). SNP microarrays are cheaper and more sensitive than aCGH for mosaicism detection. Mosaicism involving less than 5 percent variant cells has been successfully detected by this method (Conlin et al. 2010; Rodríguez-Santiago et al. 2010). NGS can provide massive parallelized genomic interrogation (Bras et al. 2012; Mardis 2008) and can detect any *de novo* sequence variants; however, the high price (around US$2000 per individual) is a barrier for screening a large number of samples.

A copy number variant (CNV) is defined as a DNA segment which is larger than 1 kb and exhibiting differences in copy number such as gains (insertions or duplications) or losses (deletions or null genotypes) (Feuk et al. 2006; Scherer et al. 2007). CNV detection based on SNP microarrays has been successfully used to explore SM in several studies (Forsberg et al. 2012; Žilina et al. 2015). The objective of this study was to explore SM of CNV between tissues, in foetuses and adult sheep, using SNP microarray technology.

5.3 Methods

5.3.1 Sample collection and genotyping

Forty-seven DNA samples were collected from 7 tissues types, epididymis, kidney, liver, semitendinosus, testis, thymus, thyroid, of 12 individuals (6 adults and 6 foetuses) (Table

5.1). The average age of adults and foetuses were 31 months and 140 days respectively. DNA was extracted from tissues using a Roche® High Pure PCR Template Preparation Kit (Roche Diagnostics GmbH, Germany). The quality of genomic DNA was assessed by agarose gel electrophoresis and quantity checked by using a Nanodrop (Thermo Fisher Scientific, USA). Forty microliters of each DNA sample, with a minimum concentration of 81.61 ng/µl, was submitted to AgResearch, Invermay Agricultural Centre, Mosgiel, NZ, for genotyping using the Ovine Infinium® HD SNP BeadChip (Rayna Anderson 2014).

5.3.2 Quality control

The Ovine Infinium® HD SNP BeadChip was designed based on the Oar_v3.1 genome assembly (http://www.livestockgenomics.csiro.au/sheep/oar3.1.php), with 606,005 SNP markers. Raw SNP output data were loaded into GenomeStudio® (V2011.1, Illumina, San Diego, CA, USA) software to extract genotype calls, error rate, GenCall score (a quality metric calculated for each genotype and ranges from 0 to 1), signal intensity (LRR) and allelic intensity (BAF) ratios for each SNP. Samples with call rate <95% were excluded. Further, quality control was done using a subprogram, filter_cnv.pl, of the PennCNV software (Wang et al. 2007). In total, 11 samples were excluded after quality control using default quality control threshold of PennCNV which is 0.3 for log R ratio standard deviation (LRRSD), 0.01 for B allele frequency (BAF) draft, 0.05 for wave factor (WF)and the remaining 36 samples were used for CNV calling (Table 5.1).

5.3.3 CNV detection

Two algorithms were used to detect CNV: PennCNV (version 1.0.3) (Wang et al. 2007) and cnvPartition v3.2.1 (Illumina, San Diego, CA, USA). SNP data pertaining to sex chromosomes was not analysed, due to the allelic imbalance in males and the reported unreasonably large variable regions on the X chromosome (Gurgul et al. 2015).

The cnvPartition plugin v3.2.0 (Illumina, San Diego, CA, USA) was installed into GenomeStudio® (v2011.1), and LRR and BAF of all SNP for 47 samples were read directly.

This algorithm employs LRR and BAF to detect systematic deviation in neighbouring SNP markers, representative of distinct copy numbers. A default threshold (confidence score of 35) was applied and the minimum number of SNP markers per CNV segment was set to three (The confidence score is defined as the sum of all logged likelihoods in the region for the assigned copy number minus the sum of all log L2 values for loci in the region, L2 is the locus with a putative copy number 2).

The final report with 36 samples information (LRR and BAF) were exported using the report wizard function of GenomeStudio®. The sub code of PennCNV, split_illumina_report.pl, was used to split final report into 36 single files of each animal for further analysis. A file containing the population frequency of B allele (PFB) of SNP was created by using the compile_pfb.pl program in PennCNV, based on SNP data from 36 Romeny sheep. The GCmodel option of PennCNV was not applied in this study because this model was not yet optimised for non-human species (Wang K, personal communication). PennCNV integrates LRR, BAF, PFB for each SNP, and the distance between adjacent SNP, into a Hidden Markov Model (HMM), for detecting CNV. The minimum number of SNP per CNV segment decides the length and resolution of CNV calling. A minimum of three SNPs per CNV segment, default setting, was assumed and CNV calling performed using the code detect_cnv.pl.

Table 5.1 Details of tissue samples analysed.

The red colour 1 represents the sample excluded, while the black colour 1 represents the samples kept based on default quality control, thresholds of PennCNV algorithm (0.3, 0.01 and 0.05 for LRRSD, BAF draft and WF, respectively).

Tissues	**Adults**						**Foetuses**						**Total/tissue**	**Total/tissue (after QC)**
	178	**255**	**317**	**464**	**503**	**716**	**18a**	**19b**	**21a**	**24b**	**26b**	**27a**		
Liver	1	1	1	1	1	1							6	4
Semitendinosus	1	1	1	1	1	1	1	1	1	1	1	1	12	10
Testis	1	1	1	1	1	1							6	6
Epididymis		1	1	1	1	1							5	1
Kidney							1	1	1	1	1	1	6	6
Thymus							1	1	1	1	1	1	6	5
Thyroid							1	1	1	1	1	1	6	4
Total/animal	3	4	4	4	4	4	4	4	4	4	4	4	**47**	
Total/animal (after QC)	2	4	3	2	3	3	3	4	3	2	3	4		**36**

The CNV outputs from the two algorithms were inputted into CNVRuler v1.5 software (Kim et al. 2012), in order to derive copy number variation range (CNVR). This programme produces CNVR by merging CNV that overlap by at least one base-pair. Derived CNVR were categorized as: 'loss' (CNVR containing deletions), 'gain' (CNVR containing duplications) and 'mixed' (CNVR containing both deletions and duplications). CNVR common between any two algorithms were determined by reciprocal overlap. Finally, the CNVR were mapped to the sheep genome using a custom written script in R (Appendix 5.1).

5.3.4 Estimation of CNV mosaicism

In order to estimate CNV mosaicism, unique CNVs in each individual were calculated by merging full CNVs detected in each individual. The number of unique CNVs represents the kinds of CNVs detected. The unique CNVs were used to do estimation of CNV mosaicism presented in Table 5.2. The number of CNV mosaicism (expressed as % mosaic CNVs) between tissues within each individual was estimated by overlapping unique CNVs derived using the cnvPartition and PennCNV algorithms, individually as well as in combination. Differences in mosaic percentages between age groups (adults versus foetuses) as well between algorithms (cnvPartition, PennCNV and combined) were tested using generalised linear mixed model analysis in SAS® 9.4 (SAS Institute Inc. Cary, NC, USA). The employed *proc mixed* model included the fixed effects of group, algorithm and their interaction, and random effects of animals.Distribution of data for mosaic percentages was tested for normality using Shapiro-Wilk, Kolmogorov-Smirnov and Anderson-Darling tests using the *univariate* procedure in SAS® 9.4. The residuals of data for mosaic percentages were found to be non-normally distributed and hence, mosaic percentages were normalised by Blom's transformation employing the *proc rank* procedure in SAS® 9.4. The option of statement, NORMAL=BLOM, was used to estimate the normal scores corresponding to the

observations as per Blom (1958). Arithmetic means and standard errors in original scale were presented in Table 5.2.

5.3.5 qPCR validation

Selected CNVs were validated by qPCR, using StepOnePlus™ Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) (Ma and Chung 2014). Four CNVs detected in eight sheep were validated. The *DGAT1* gene was used as reference since it was shown to be free from copy number variation (Fontanesi et al. 2011). Details of primer sequences, target regions in the sheep map, as well as PCR conditions are shown in additional files: Table S5.2 qPCRresult. In total, 12 samples (5 foetuses and 7 adults) from six different tissues (thymus, kidney, semitendinosus, liver, testis and epididymis) were tested.

The copy number of the amplified regions was calculated by a relative standard curve method (Biosystems 2004) as follow:

$Qty = 10^{\frac{Ct-b}{m}}$, where *Qty*, *Ct*, m and b are the relative quantity of amplified fragment, threshold cycle, slope and y-intercept of the standard curve.

$$copy\ \ number = \frac{Qty(\text{NormalizedTarget})}{Qty(\text{NormalizedReference})} = \frac{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{target sample}}{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{reference sample}}$$

Table 5.2 Summary of CNV mosaicism detected in foetal and adult sheep, using cnvPartition and PennCNV alone, or in combination.

Age group	Animal ID	Number of tissues genotyped	Algorithm used								
			cnvPartition			PennCNV			Both cnvPartition and PennCNV		
			Total number of unique CNV	Number of mosaic CNV	Mosaic CNV percentage	Total number of unique CNV	Number of mosaic CNV	Mosaic CNV percentage	Total number of unique CNV	Number of mosaic CNV	Mosaic CNV percentage
Adults	178	2	50	28	56.0%	44	22	50.0%	14	2	14.3%
	255	4	73	61	83.6%	41	28	68.3%	18	10	55.6%
	317	3	77	54	70.1%	84	64	76.2%	27	12	44.4%
	464	2	59	30	50.8%	37	16	43.2%	13	3	23.1%
	503	3	90	64	71.1%	105	78	74.3%	35	17	48.6%
	716	3	98	77	78.6%	136	114	83.8%	43	30	69.8%
	Overall				68.4^{ao}%			66.0^{ao}%			42.6^{bo}%
Foetuses	18a	3	75	53	70.7%	63	40	63.5%	26	15	57.7%
	19b	4	65	54	83.1%	163	150	92.0%	17	9	52.9%
	21a	3	126	113	89.7%	186	172	92.5%	35	26	74.3%
	24b	2	51	20	39.2%	63	30	47.6%	27	10	37.0%
	26b	3	97	83	85.6%	129	117	90.7%	31	22	71.0%
	27a	4	78	56	71.8%	136	113	83.1%	31	18	58.1%
	Overall				73.3^{ao}%			78.2^{ao}%			58.5^{bo}%

Note: Overall mosaic percentages with different superscripts, within each group (adults and foetuses), differ significantly ($P<0.05$)

However, it is difficult to find a standard sample as a reference which has copy number variation. Therefore, firstly, the reference was selected randomly. The copy number of reference was assumed as 1 copy, then calculate the accuracy of qPCR. After that, the copy number of reference was assumed as 2, and 3 and did the same process again. Because the gene has more than 4 copies is rare, no more assumption was set up. By comparing the accuracy between the copy numbers 1, 2, 3 the copy number which has highest correction rate is considered as the correct copy number. Of course, this method could have bias. since the assumption of copy number in reference could be wrong. The thresholds table is given below (Table 5.3).

Table 5.3 Hypothetical copy numbers in the reference sample and their thresholds (based on qPCR) for copy number evaluation

hypothetical copy number of the control	**1 copy**	**2 copies**	**3 copies**	**4 copies**
1 copy	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5
2 copies	0.25-0.75	0.75-1.25	1.25-1.75	1.75-2.25
3 copies	0-0.459	0.459-0.825	0.825-1.165	1.165-1.495

5.3.6 Gene annotation

By using a custom written Perl and MySQL script written by A/Prof Patrick Biggs, position information of CNVs was matched to Oar_v3.1 assembly in order to find genes (from the Ensembl database) located in the detected CNVs. Functional annotations of identified genes were made using the online tool, bioDBnet (biodbnet-abcc.ncifcrf.gov/db/db2db.php).

5.4 Results

5.4.1 CNV detection and CNVR formation

The Ovine Infinium® HD SNP BeadChip was used to detect CNV in 36 tissue samples obtained from 12 individual sheep. In total, 1,546 and 1,764 CNV segments were detected using cnvPartition and PennCNV algorithms, respectively (Additional file Table S5.1). Of them, 664 CNVs were detected by both algorithms and hence, these CNV could be considered more reliable (Table 5.4). CNVs detected by both algorithms, that overlapped each other, were merged and a total of 140 CNVRs were formed (Figure 5.1; Additional file Table S5.1), ranging between 318 bp and 450 kb in size.

There was little between-individual overlap of the detected CNV, within the two age groups. In adults, only one CNV of 77 unique CNVs was shared by all six individuals (Figure 5.2), while in foetuses, there was no CNV of 43 unique CNVs found in all six individuals and only one CNV found in five individuals (Figure 5.3).

Table 5.4 Summary of CNVs detected by cnvPartition, PennCNV and their combination.

cnvPartition			**PennCNV**			**Both algorithms**		
Loss	Gain	Total	Loss	Gain	Total	Loss	Gain	Total
1,251	295	1,546	1,275	489	1,764	442	222	664

5.4.2 CNVR differences between adults and foetuses

Comparison of CNVR between adults and foetuses revealed that only 45 out of total 140 (32.1 %) CNVR were common to both groups (Figure 5.4, Figure 5.5). The total number of CNVR and average nmber of CNVR per sample were almost similar in the two age groups (Table 5.5).

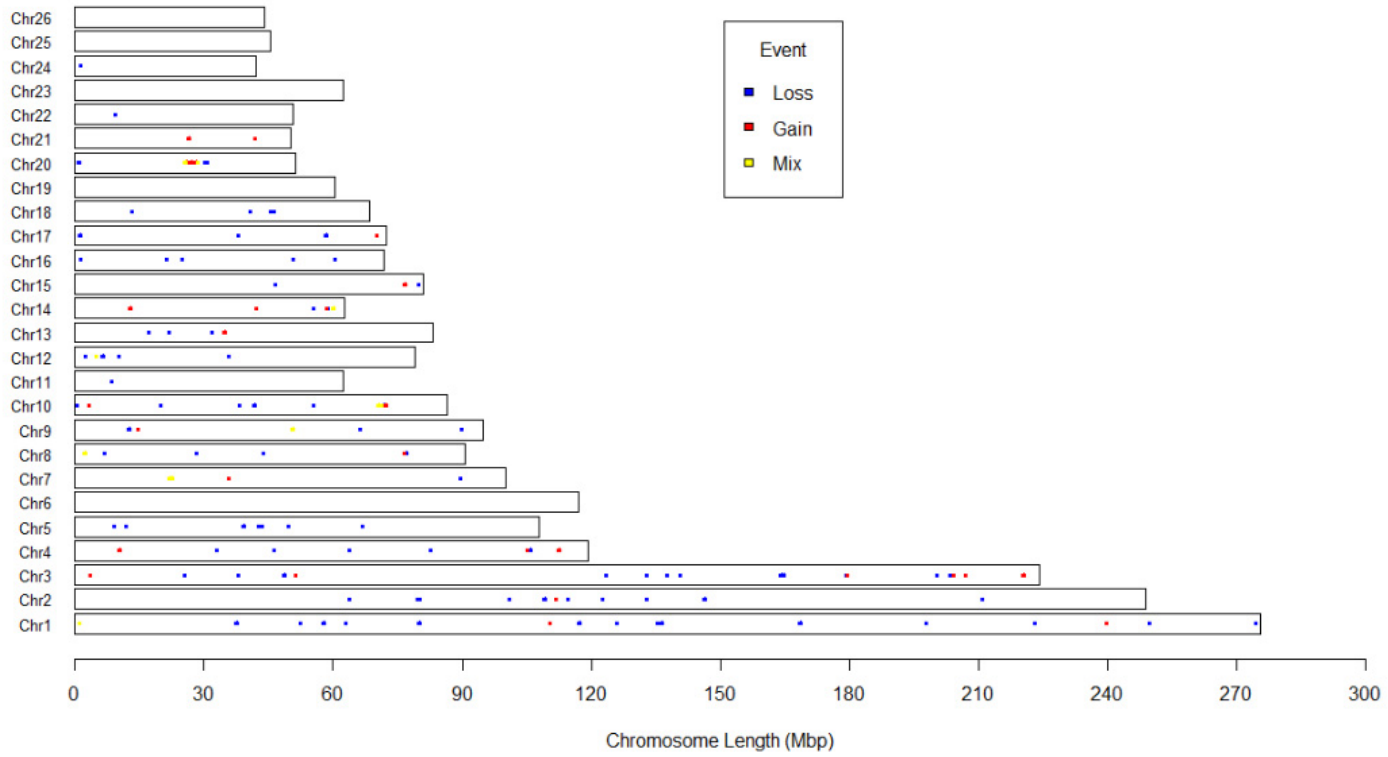


Figure 5.1 Distribution map of CNVRs detected by both PennCNV and cnvPartition using 36 tissues from 12 sheep (adults and fetuses both).

The x axis denotes position across chromosomes and dots in different colours (blue represents losses, red gains and yellow mixed) denote location of CNVRs.

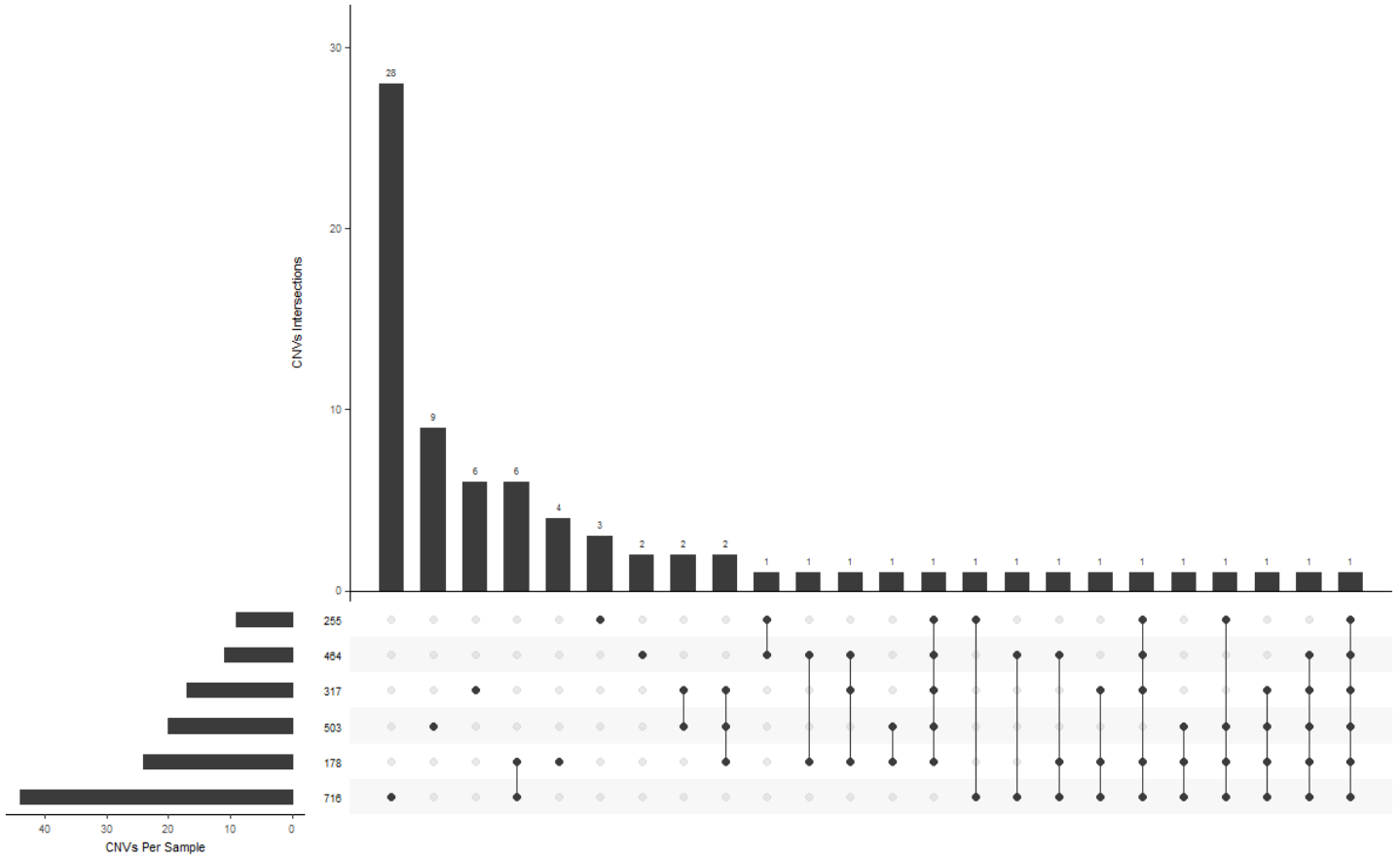


Figure 5.2 UpsetR plot showing overlap of CNVs in six adult sheep.

The y-axis represents the number of CNVs, while black circles with connecting lines beneath the x-axis denote sharing of respective CNVs between the different individuals. The horizontal bars at the bottom left corner depict the number of CNVs per sample in each individual.

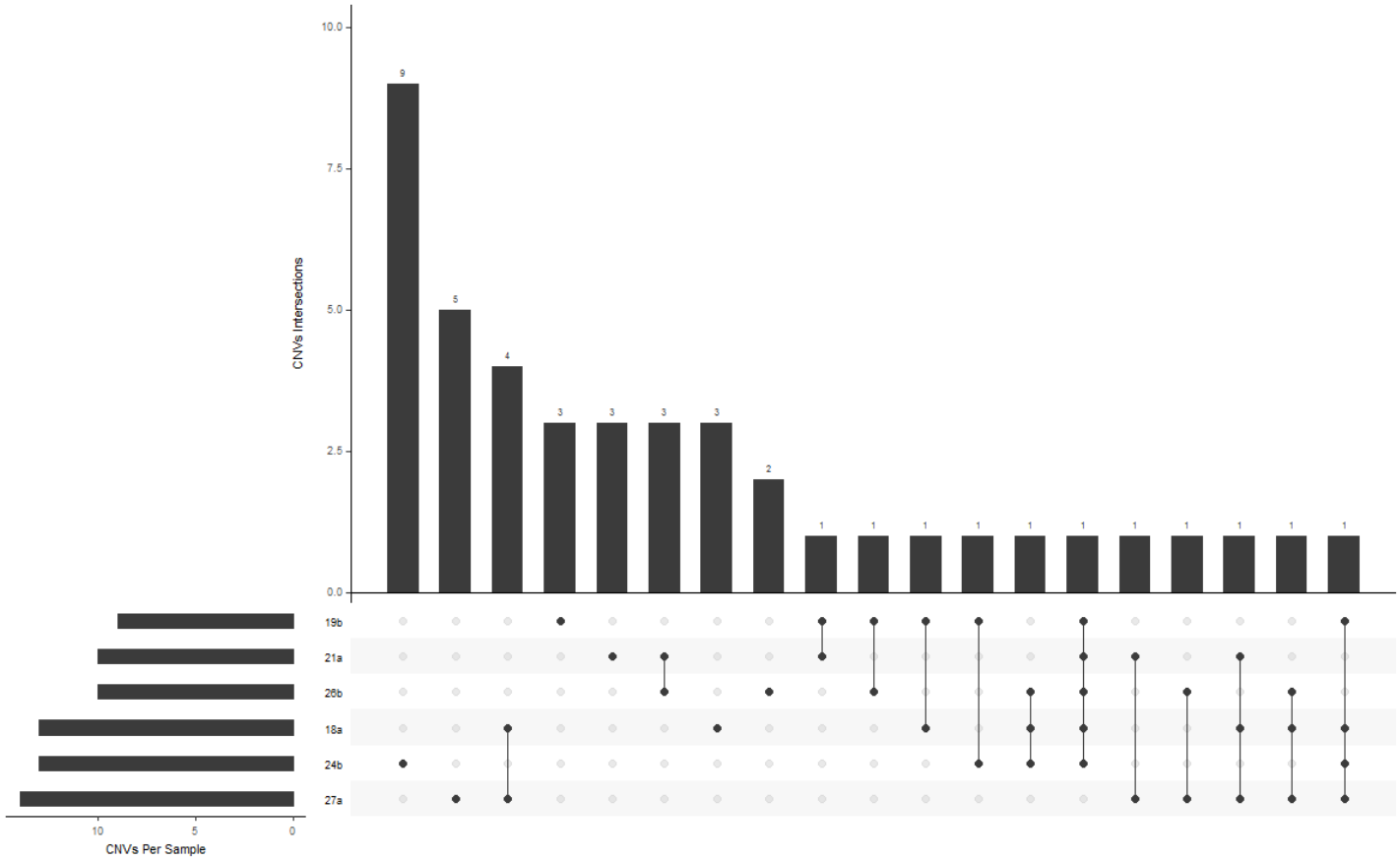


Figure 5.3 UpsetR plot showing overlap of CNVs in six foetuses.

The y-axis represents the number of CNVs, while black circles with connecting lines beneath the x-axis denote sharing of respective CNVs between the different individuals. The horizontal bars at the bottom left corner depict the number of CNVs per sample in each individual.

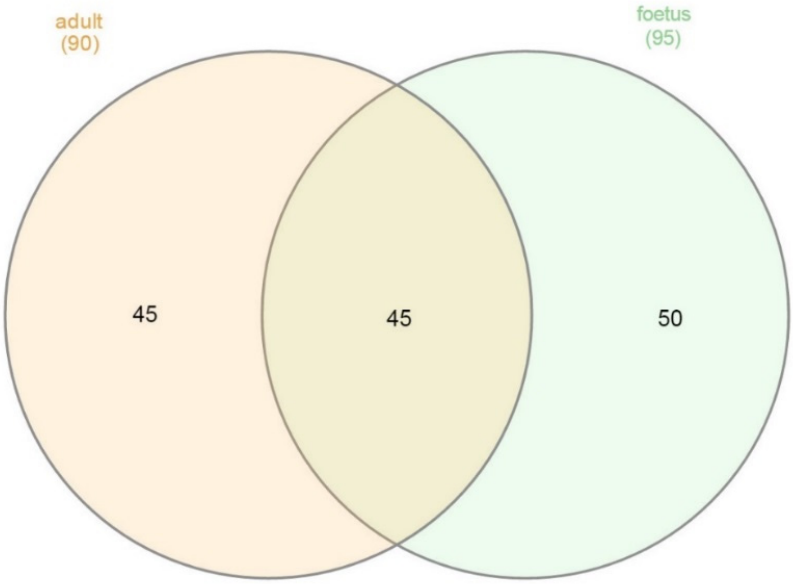


Figure 5.4 Venn plot of CNVR detected in tissues from adults and foetuses.

The pink and green circles represent CNVR in adults and foetuses, respectively. The numbers in overlapping regions denote the number of CNVR common to both groups while those in non-overlapping regions are unique for each group.

Table 5.5 CNVR differences between adults and foetuses.

	CNVR	**Samples**	**CNVR/sample**
Adults	90	17	5.29
Foetuses	95	19	5
Overall	185	36	5.13

5.4.3 CNVR differences between tissues

CNVRs were formed by merging CNVs of each tissue from all samples and plotted (Figure 5.6-Figure 5.8) using custom written code in R (Appendix 5.2).

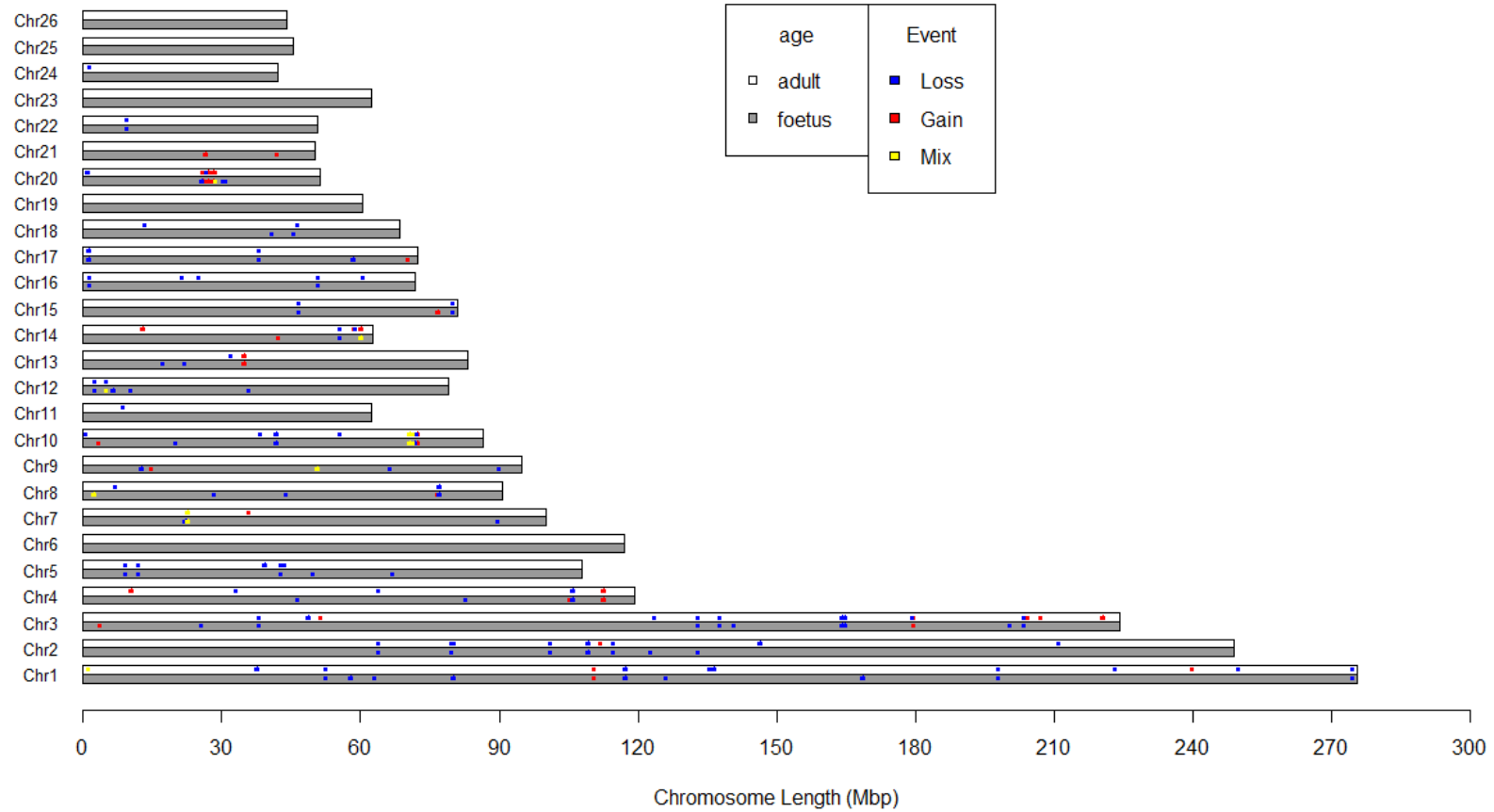


Figure 5.5 Distribution of difference types of CNVRs in individual chromosomes of adults and foetuses.

The x-axis denotes position across chromosomes and dots in different colours (blue represents losses, red gains and yellow mixed) denote location of CNVRs on chromosomes in adults (white bars) and foetuses (grey bars).

Of the total 140 CNVRs that were detected, only six CNVRs were found in all tissues, while 55 were tissue-specific (Figure 5.9). The total number of CNVRs as well as the average number of CNVRs per sample varied between tissues (Table 5.6). The average number of CNVR per sample was highest in the liver (17.75) and lowest in the semitendinosus (7.50).

Table 5.6 Number of CNVRs detected in individual tissues across animals

Tissue	Number of samples	Number of CNVRs detected	CNVR/sample
Epididymis	1	13	13.00
Kidney	6	61	10.17
Liver	4	71	17.75
Semitendinosus	10	75	7.50
Testis	6	58	9.67
Thymus	5	51	10.20
Thyroid	4	60	15.00

5.4.4 Within-individual SM of CNV

CNV mosaicism in foetal and adult sheep was significantly (P<0.05) higher when either cnvPartition or PennCNV algorithms were used for CNV detection, compared to their combined usage (Table 5.2). In foetuses, the overall percentage mosaicism identified by cnvPartition, PennCNV and the combination were 73.3, 78.2 and 58.5% respectively. The respective percentages in adult sheep were 68.4, 66.0 and 42.6%. There was no significant difference (Table 5.2) in percent mosaicism detected by the two individual algorithms, in either adults or foetuses. Irrespective of age group and algorithm used, mosaicism, tended to be higher in individuals which had a greater number of tissues investigated. That is, animals with either four or three tissues examined exhibited higher mosaicism, compared to those with just two tissues. Although foetuses exhibited higher mosaicism, compared to adults, irrespective of algorithm used, the difference was non-significant (Table 5.2).

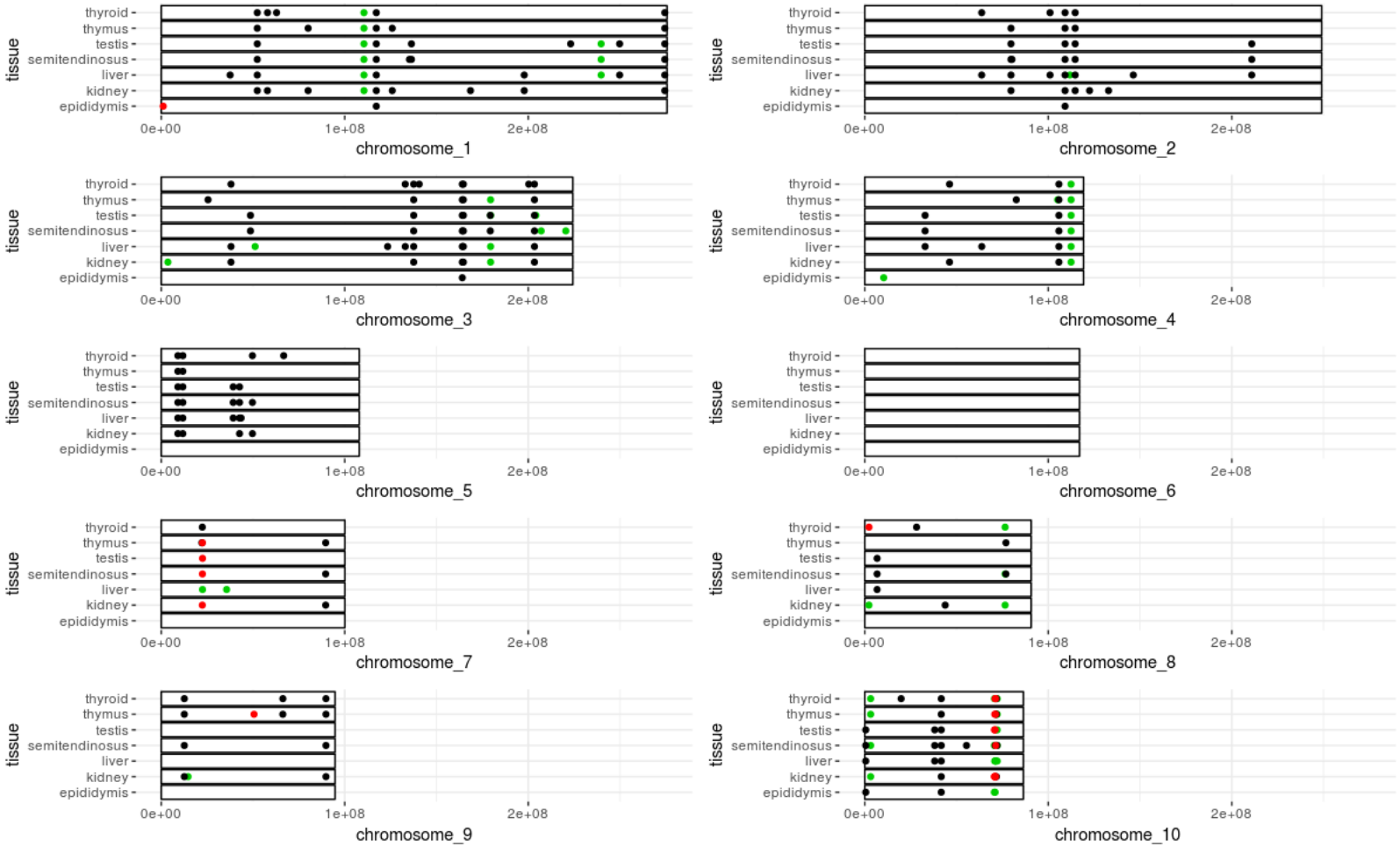


Figure 5.6 Distribution map of CNVRs in different tissues from chromosome 1 to chromosome 10.

The x-axis denotes position across chromosomes and dots in different colours (black represents losses, red gains and green mixed) denote location of CNVRs on chromosomes in different tissues (depicted in seven grading colours, from white to dark grey, for epididymis, kidney, liver, semitendinosus, testis, thymus, thyroid, respectively).

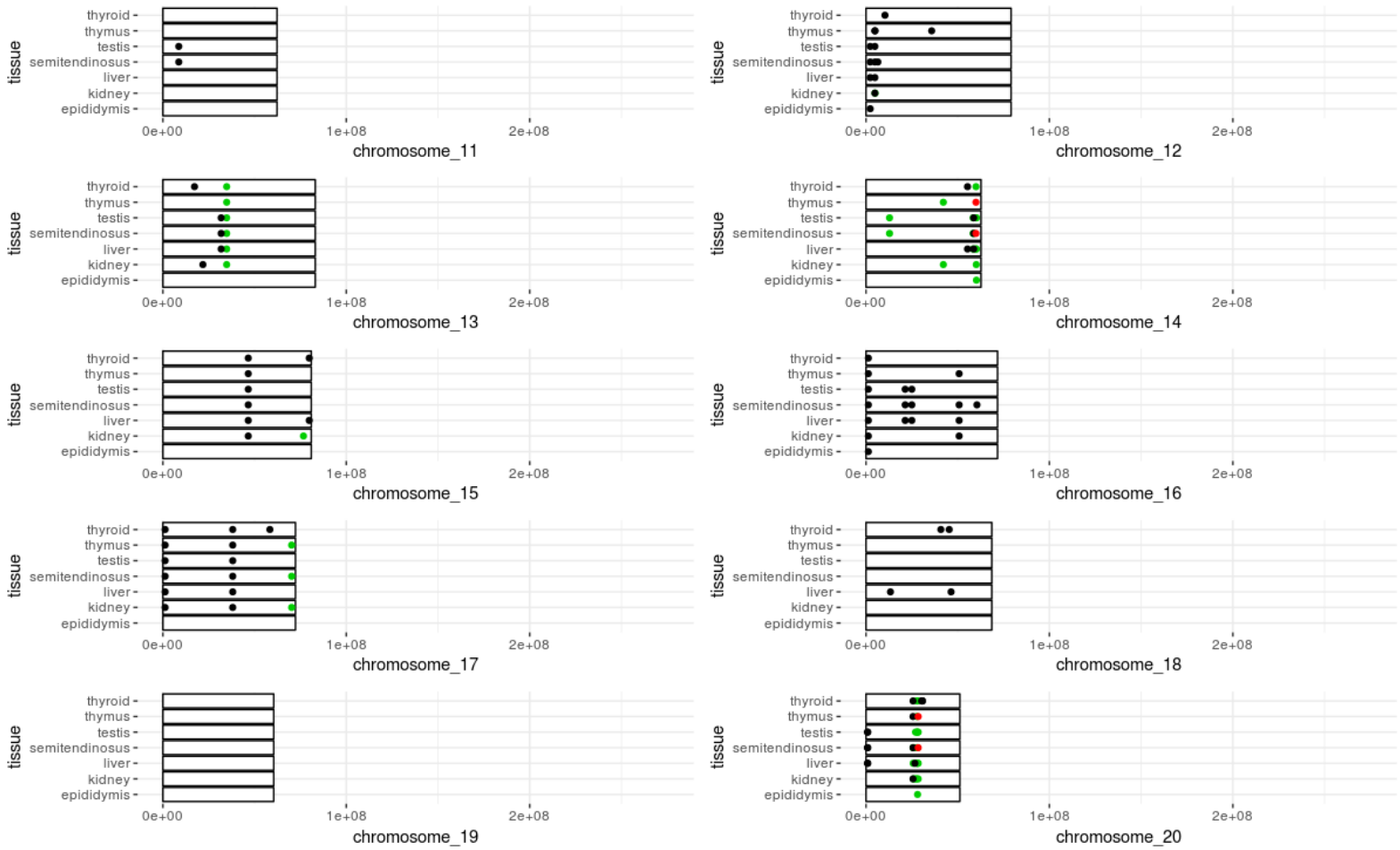


Figure 5.7 Distribution map of CNVRs in different tissues from chromosome 11 to chromosome 20.

The x-axis denotes position across chromosomes and dots in different colours (black represents losses, red gains and green mixed) denote location of CNVRs on chromosomes in different tissues (depicted in seven grading colours, from white to dark grey, for epididymis, kidney, liver, semitendinosus, testis, thymus, thyroid, respectively).

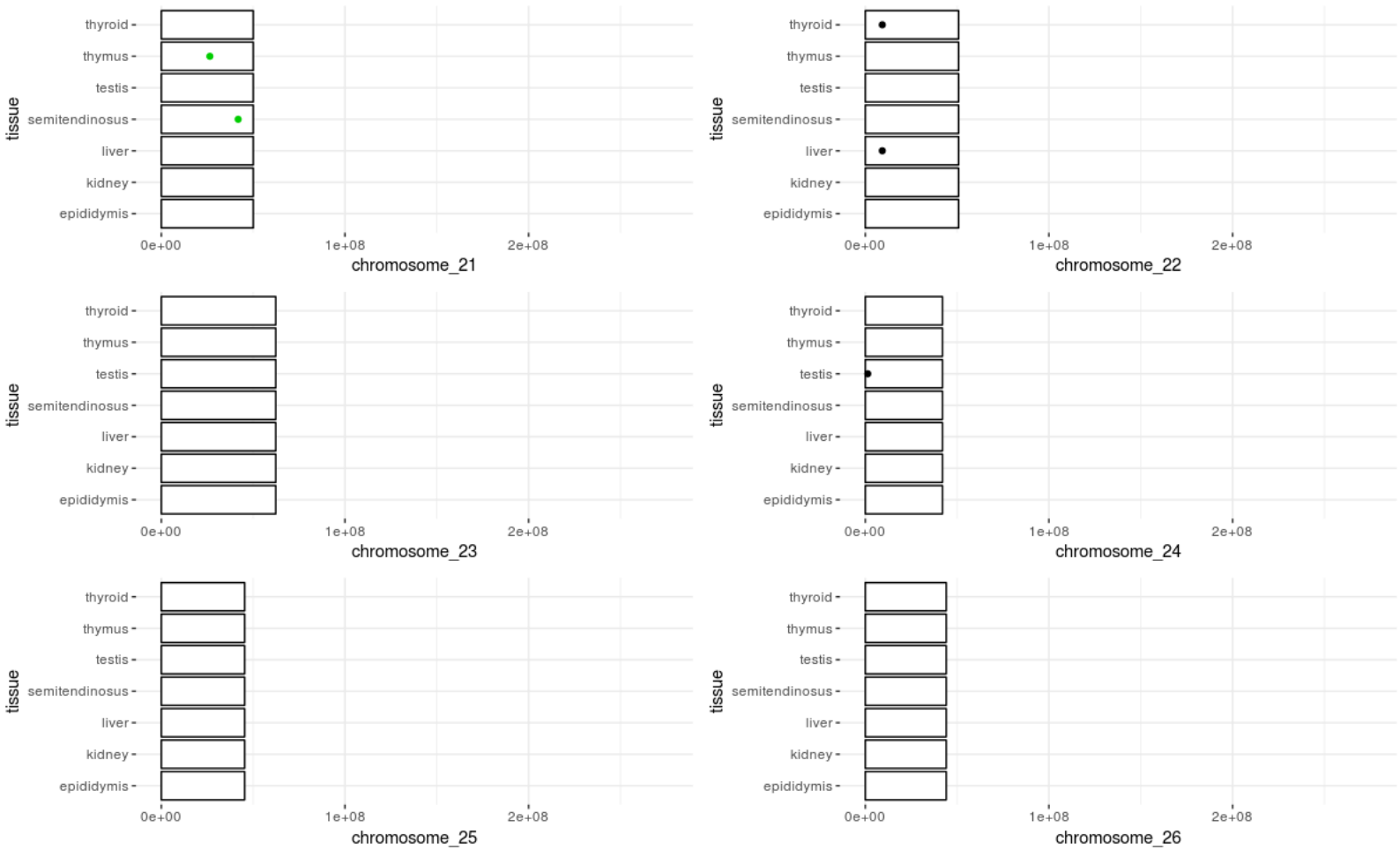


Figure 5.8 Distribution map of CNVRs in different tissues from chromosome 21 to chromosome 26.

The x-axis denotes position across chromosomes and dots in different colours (black represents losses, red gains and green mixed) denote location of CNVRs on chromosomes in different tissues (depicted in seven grading colours, from white to dark grey, for epididymis, kidney, liver, semitendinosus, testis, thymus, thyroid, respectively).

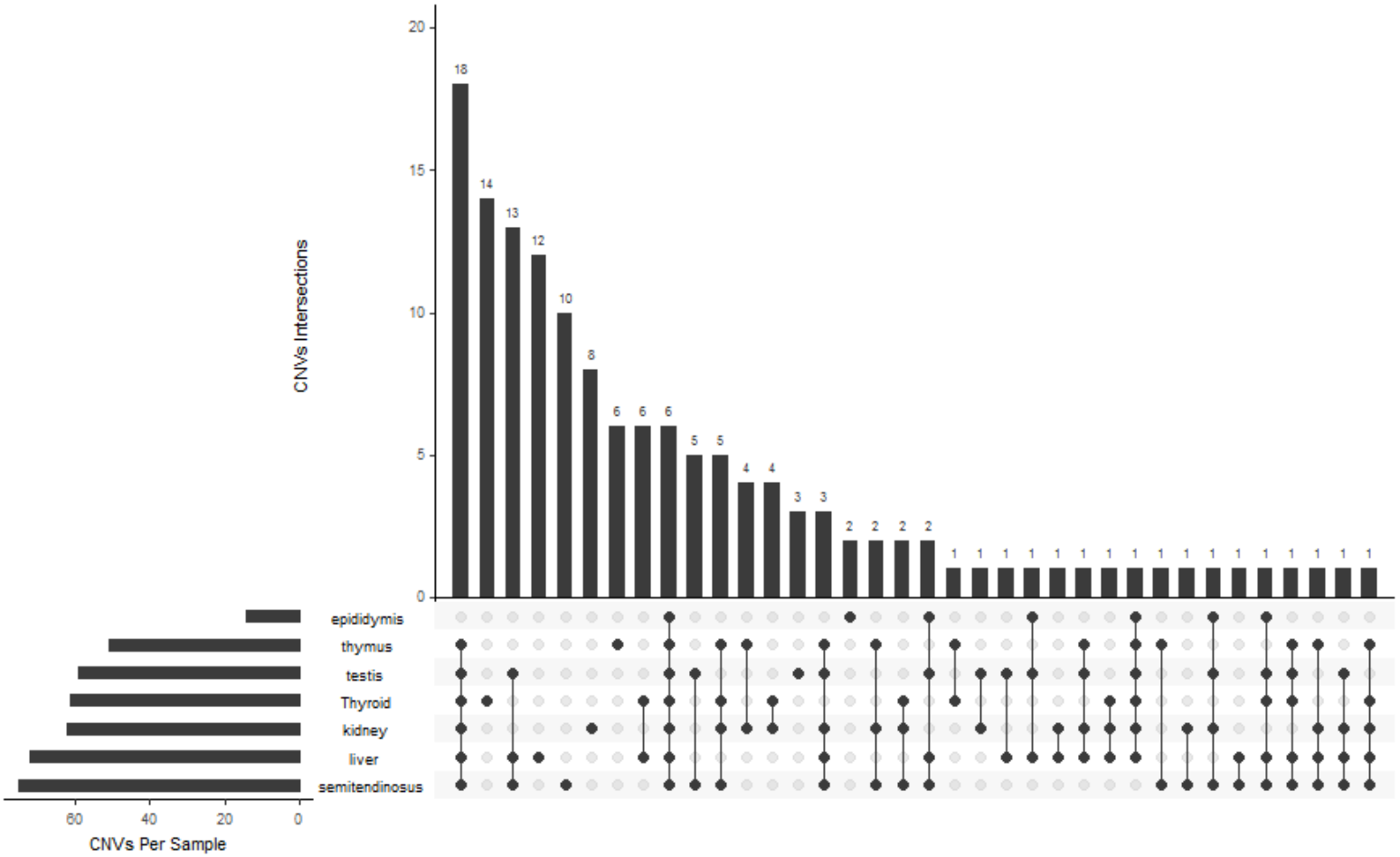


Figure 5.9 UpsetR plot showing overlap of CNVRs across seven tissues made by UPSetR (Conway et al. 2017).

Y-axis represents the number of CNV, while black circles with connecting lines beneath the X-axis denote sharing of respective CNVs between the different tissues. The horizontal bars at the bottom left corner depict number of CNV per sample in each tissue.

5.4.5 qPCR validation

Four randomly selected CNV segments were validated by qPCR, using DNA from 12 samples. Validation accuracy with regard to the four markers ranged between 58.3% to 75.0%, with an overall mean of 66.7% (Table 5.6; Additional file: S5.2)

Table 5.6 Results of qPCR validation four randomly selected CNVs detected by both the cnvPartition and PennCNV algorithms

Primer ID	Samples validated	Validation accuracy	
		Proportion	Percentage
J14	12	7/12	58.3%
J16	12	8/12	66.7%
J20	12	8/12	66.7%
J22	12	9/12	75.0%
Total	48	32/48	66.7%

5.4.6 Gene annotation

The majority of the CNVs detected in this study were located in non-coding regions of the genome. Only 17 genes (Additional file S5.1), such as mitochondrial translational release factor 1, 5'-nucleotidase, HEAT repeat containing 5A, ubiquitin C-terminal hydrolase L5, were found to be present in the detected CNV.

5.5 Discussion

Somatic mosaicism of CNV has been reported in humans (O'Huallachain et al. 2012; Žilina et al. 2015) and animals (Jung et al. 2013). In order to detect CNV mosaicism in sheep, 47 tissue samples (from seven different tissues) were obtained from six foetuses and six adult sheep and genotyped using the Ovine Infinium® HD SNP BeadChip. Two algorithms, PennCNV and cnvPartition, either alone or in combination were used to detect CNV from the resulting SNP genotype data.

There were differences in CNV detection between algorithms. Overall, 693 and 944 somatic mosaicism CNVs were detected using cnvPartition and PennCNV, respectively, which represented 73.8% and 79.5% of total CNVs detected by respective algorithms. By considering only the CNV detected by both algorithms (highly reliable CNV), 174 out of the total 317 (54.9%) CNVs exhibited somatic mosaicism.

A large difference in CNVs was found between individuals. The study results showed that only one CNV was common in all adult sheep and 67.5% of CNV detected by both algorithms were unique CNVs i.e. existed in only one individual (Figure 5.2). No CNV was detected in all six foetuses and the percentage of unique CNVs was 58.1% (Figure 5.3). This indicates that the formation of CNV could be a random event such as mutation. Also, a few CNVs are shared by a few individuals and this could be due to inheritance of those CNVs from a common ancestor because samples were collected from the same population.

There were differences in CNVRs between the age groups. Only 45 CNVRs were shared between adults and fetuses. Forsberg et al (2012) indicated that age could be a factor influencing the formation of CNV in human blood cells (Forsberg et al. 2012), the observed CNVRs differences between age groups, in the current study, could just be reflective of diversity between individuals, due to a limitation in the current study design (where samples from adult and foetal tissues were obtained from different individuals) so that for further study it is necessary to collect sample from identical twins. CNVRs differences also existed between tissues. More than one third of the total CNVRs detected (55 out of 140) were unique to different tissues and the number of CNVRs/sample in seven tissues varied between 7.50 and 17.75 (Table 5.6), indicating that there might be differential mechanisms in the tissues contributing to CNV formation during the development of the embryo. The small and uneven sample size (across tissues) could be another contributing factor for this fluctuation.

Analysis of SM, in the current study, showed interesting results. The percentage of mosaic CNVs in adults as well as foetuses, using either cnvPartition or PennCNV alone were significantly higher compared to those found when they were used in combination (Table 5.2). This indicates increased reliability of CNV prediction when the two algorithms were used in combination. A similar approach, namely a combination of PennCNV and QuantiSNP algorithms, was used to investigate SM in humans (Zilina et al. 2015). However, compared with previous studies, this study found many more mosaic CNVs. Zilina et al (Žilina et al. 2015) found 3 mosaic CNVRs (16.6%) in human and Seung-Hyun et al (Jung et al. 2013) found 5 mosaic CNVs (6%) in dogs which indicated samples, platform and algorithms could bring huge differences to CNV detection (Žilina et al. 2015).

Also, in the current study, individuals with three or four tissues genotyped exhibited higher SM, compared to those with only two tissues analyzed, irrespective of the algorithm employed (Table 5.2 and Figure 5.10), indicating a correlation between mosaicism and number of tissues analyzed. In addition, the average mosaic percentage in foetuses is higher (although non-significant) than that in adults, which could be because of difference in the tissue types in the two groups.

The accuracy of CNV prediction in the current study was verified using qPCR. Results showed a validation accuracy of 66.7%, which is slightly lower than that found in a CNV study in cattle (75%), using NGS data (Boussaha et al. 2015) and higher (43.4%) than that in a different study in humans, that employed the Nimblegen 2.1M oligonucleotide whole-genome array (O'Huallachain et al. 2012). This indicates that employing a combination of two algorithms is adequate to achieve a reasonable accuracy of CNV prediction on the Illumina SNP platform.

Gene annotation showed that most of the detected CNVRs were located in the non-coding regions. Only 17 genes were found to be located within the CNVRs. This is consistent with studies in humans (Žilina et al. 2015) and animals (Jung et al. 2013) which also found the majority of CNVs to be located in non-coding regions of the genome.

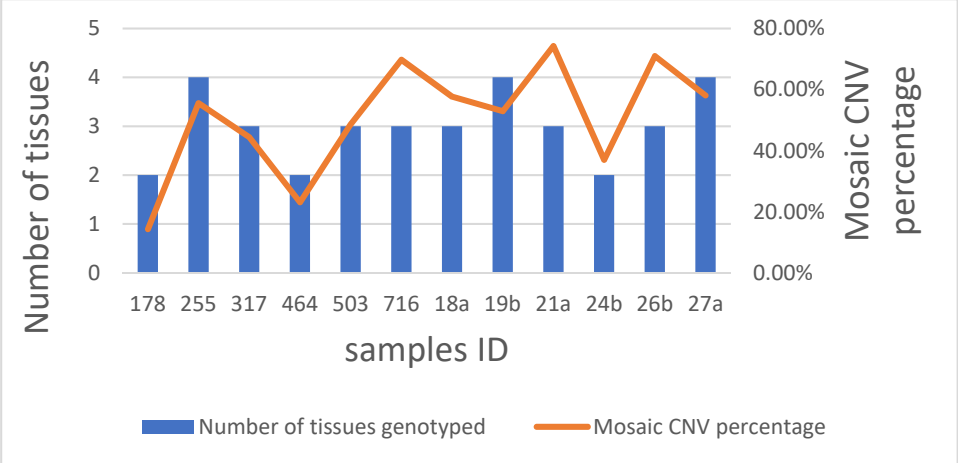


Figure 5.10 Relationship between CNV mosaicism and number of tissues investigated.

CNVs detected by both cnvPartition and PennCNV were used to estimate per cent mosaicism (shown in red line and scaled to right-side Y axis). Blue bars represent number of tissues genotyped (scaled to left-side Y axis).in each individual (shown along X axis).

5.6 Conclusion

This study showed that a high degree of mosaicism of CNV was found to exist in sheep and it could be influenced by age, individuals, CNV detection algorithm as well as tissues analysed. Employing a combination of CNV detection algorithms, rather than individual algorithms is crucial in order to achieve a reasonably high accuracy to estimate SM.

5.7 Additional files

Additional file: Table S5.1

Chapter 6

Detection of copy number variation and genome-wide positive selection signatures using Ovine Infinium® HD SNP BeadChip in two Romney lines, selected for resistance or resilience to gastrointestinal nematodes

Juncong Yan[1], Hugh T. Blair[1], Andrew Greer[2], Joseph Hamie[2], Patrick Biggs[1], Venkata S.R. Dukkipati[1*]

To be submitted to BMC Genomics

[1] IVABS, Massey University, Palmerston North 4442, New Zealand
[2] Agricultural and Life Sciences, Lincoln University, Lincoln 7647, New Zealand
[*] Correspondence: R.Dukkipati@massey.ac.nz

6.1 Abstract

6.1.1 Background

Gastrointestinal nematodes are one of the most serious parasitic threats for sheep and the large-scale usage of deworming drugs is not welcomed by the present global market owing to anthelmintic resistance. Therefore, a new anti-parasite strategy is necessary. Genetic breeding is one of the most important ways in animal husbandry to improve the quality of domestic animal. This study was undertaken to explore differences in copy number variation (CNV) and detect single-nucleotide polymorphism (SNP) based selection signatures in two Romney sheep lines, selectively bred for nematode resilience or resistance.

6.1.2 Result

Ninety-three sheep from the two selection lines were genotyped using the Ovine Infinium® HD SNP BeadChip, and extended haplotype homozygosity (EHH) and site-specific extended haplotype homozygosity (EHHS) analyses were undertaken to detect selection signatures. Also, copy number variation (CNV) was investigated in the two lines. In total, 224 SNPs (147 in EHH and 77 in EHHS), harboured within 45 genes (36 in EHH and 9 in EHHS), were found to be significant ($P<0.0001$). Ten SNPs found by both XP-EHH and Rsb were located within two previously identified QTLs, LATRICH_2 and FECGEN, associated with nematode larval count and faecal egg count, respectively.

There were 1,805 and 1,508 CNVs detected in the resilient and resistant lines, respectively, which formed 314 (137 gains, 145losses and 32 mixes) and 293 (112 gains, 153 losses and 28 mixes) CNVRs respectively. Only 196 CNVRs were common between the two lines. Besides, 117 genes (69 and 48 in the resilient and resistant lines, respectively) were found in the CNVRs that were unique to the two lines. Also, the CNVRs overlapped with 49 known parasite related QTL zones.

None of the SNPs that were significant in overlapped results of EHHS test between Rsb and XP-EHH are located on CNVRs of resilience or resistance, but one SNP that was significant in Rsb only was located within a unique CNVR detected in the resilient line.

6.1.3 Conclusion

This study provided a genome-wide map of positive selection signatures and CNV detection in two Romney sheep lines selected for gastrointestinal nematode FEC. Dozens of significant SNPs were identified and ten of them were found to be located within two previously detected QTLs associated with gastrointestinal nematodiasis in sheep. Huge between-line CNV differences were evident. Except for one, none of the significant SNPs overlapped to the detected CNV regions, indicating SNP-based selection signatures and CNV could represent different aspects of genetics of sheep resistance\resilience to nematodes.

6.1.4 **Keywords:** sheep, positive selection signature, nematodes

6.2 Background

Gastrointestinal nematodes are one of the most serious parasitic threats for sheep (Familton and McAnulty 1997; Perry and Randolph 1999), costing approximately $300 million annually to the New Zealand sheep industry (Rattray 2003). The current use of anthelmintics is not welcomed by the present global market, owing to a combination of the development of anthelmintic resistance and increasing consumer preference for chemical-free food products. Therefore, alternative anti-parasite strategies are necessary. Genetic breeding is one of the most important ways in animal husbandry to improve the quality of domestic animals. Several studies have shown that resistance to nematodiasis in sheep is highly variable and heritable between individuals so that selective breeding can be an alternative and successful choice for nematode control (Morris et al. 1995; Morris et al. 2000; Morris et al. 2005).

The advent of high-density single nucleotide polymorphism (SNP) chips has facilitated detection of copy number variation (CNV) and artificial selection signatures based on patterns of linkage disequilibrium in selection lines.

Copy number variants (CNV) are defined as segments of DNA (larger than 1 kb) displaying copy number differences such as gains (insertions or duplications) or losses (deletions or null genotypes) (Feuk et al. 2006; Scherer et al. 2007). CNV is a different kind of genetic variation from single nucleotide polymorphism (SNP) and has been shown to contribute to genetic variation in production and disease traits.

Selection signature is another useful method to detect genetic markers associated with traits, and is based on the assumption that the frequency of a novel mutation consequent to positive selection will increase more rapidly than that of a neutral mutation (Sabeti et al. 2002). Consequently, long linkage disequilibrium (LD) blocks involving the mutant genes could exist in lines undergoing artificial selection since there would not be enough generations to break the LD by recombination (Slatkin 2008). Hence, a high frequency and unusually long haplotype could indicate the presence of a positive selection signature. Selection signatures based on SNP have been widely used in sheep (Gouveia et al. 2014; McRae et al. 2014) as well as cattle (Stella et al.2010; Bahbahani et al. 2018) studies.

To detect selection signatures, an algorithm called extended haplotype homozygosity (EHH) was initially introduced (Sabeti et al. 2002) and this quantifies the decay of haplotype homozygosity within a family line. Subsequently, another method known as the site-specific extended haplotype homozygosity (EHHS), was introduced to do the same purpose between family lines (Sabeti et al. 2007). These methods have been successfully used to detect selection signatures in animals (McRae et al. 2014; Somavilla et al. 2014; Zhang et al. 2012). The objectives of this study were: 1) explore CNV in two Romney sheep lines selected for

either resistance or resilience to gastro-intestinal parasite infections and 2) detect selection signatures using SNP haplotype data.

6.3 Materials and Methods

6.3.1 Ethics statement

This study was carried out following the guideline of the 1999 New Zealand Animal Welfare Act and was approved by Lincoln University Animal Ethics Committee (Permit Number: LUAEC#588)

6.3.2 Sample collection and background of lines

Ear punch samples, using an Allflex tissue sampling unit, (Allflex New Zealand Ltd, Palmerston North, New Zealand) were obtained from 93 Romney sheep belonging to two selection lines (nematode resistant, n = 42, and nematode resilient, n = 51), currently being maintained at Lincoln University, New Zealand. These two lines were selected from 1985 to 2009 (24 years), based on faecal egg count (FEC) using best linear unbiased prediction (BLUP) techniques. Since 2010, sheep within each line were randomly mated. Details regarding the selection lines were described elsewhere (Baker et al. 1990). In this paper, resilient equal susceptible in other papers published by Agresearch. The tissue samples were submitted to AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand, for DNA extraction and SNP genotyping using the Ovine Infinium® HD SNP BeadChip.

6.3.3 Quality control and data preparation for selective signature

The original SNP data (idat files) was converted to ped and map file from GenomeStudio® using PLINK Input Report Plug-in v2.1.4. (Illumina, San Diego CA, USA). Quality control was performed using PLINK_v1.9 (Chang et al. 2015; Purcell et al. 2007). A within individual call rate threshold of 99% was applied and SNPs with a call rate <95% or minor allele frequency <1%, or p value of $<10^{-6}$ for Hardy-Weinberg equilibrium were excluded.

This study considered only the SNPs located on autosomes, because the SNP probes of X and Y chromosome are not similarly distributed.

The filtered SNP data was inputted into fastPHASE_v1.4 (Scheet and Stephens 2006) in order to reconstruct the haplotypes for each autosome, using the default parameters (Appendix 6.1). The ancestral allele was determined by a simulation of fastPHASE_v1.4. The resultant haplotype data was used to detect positive selection signatures.

6.3.4 CNV detection and validation

The Ovine Infinium® HD SNP BeadChip was designed based on Oar_v3.1 gene map. The original SNP data (idat files) was converted to ped and map file from GenomeStudio® using PLINK Input Report Plug-in v2.1.4. Then, quality control of SNP was done based on call rate of samples, Minor Allele Frequency (MAF) and Call Frequency (Call Freq) using PLINK1.9 beta. SNPs with call rate less than 99% or MAF less than 1% were excluded. Individual sample was excluded if the overall SNP call rate was less than 97%. Additionally, SNPs that were not in Hardy-Weinberg equilibrium (HWE; $p<10^{-6}$) were also excluded.

CNVs were detected using PennCNV v1.03 at 3 minSNPs (Wang et al. 2007). PennCNV employs an integrated hidden Markov model on an Illumina platform. Based on three criteria, population frequency of B allele (PFB), SNP genome coordinates and a trained hidden Markov model (HMM) file, the most likely state-transition path could be analysed using the Viterbi algorithm. A PFB file of SNPs was created using the compile_pfb.pl program in PennCNV, based on SNP data from all 93 samples of this study. The GCmodel option of PennCNV was not applied in this study because this model was not yet optimised for non-human species (Wang K, personal communication). Signal intensity files, which had Log R ratio (LRR) and B Allele Frequency (BAF), were created from the final report from GenomeStudio® using a PennCNV plugin, split_illumina_report.pl. PennCNV integrates

LRR, BAF and PFB for each SNP, and the distance between adjacent SNPs, into a HMM, for detecting CNV. Final quality control of the detected CNVs was done using a program, filter_cnv.pl , of PennCNV software (Wang et al. 2007). The CNV output was inputted into CNVRuler v1.5 software (Kim et al. 2012), in order to derive copy number variation range (CNVR). This programme produces CNVR by merging CNV that overlap by at least one base-pair. Derived CNVR were categorised as: 'loss' (CNVR containing deletions), 'gain' (CNVR containing duplications) and 'mixed' (CNVR containing both deletions and duplications).

Four selected CNVs in eight individuals were validated by qPCR, using StepOnePlus™ Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) (Ma and Chung 2014). The *DGAT1* gene was used as reference since it was shown to be free from copy number variation (Fontanesi et al. 2011). Details of primer sequences, target regions in the sheep map, as well as PCR conditions are shown in Additional files: Table S6.8 qPCRresult.

The copy number of the amplified regions was calculated by a relative standard curve method (Biosystems 2004) as follow:

$Qty = 10^{\frac{Ct-b}{m}}$, where Qty, Ct, m and b are the relative quantity of amplified fragment, threshold cycle, slope and y-intercept of the standard curve.

$$copy\ \ number = \frac{Qty(\text{NormalizedTarget})}{Qty(\text{NormalizedReference})} = \frac{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{target sample}}{\left(\frac{Qty\text{Target}}{Qty\text{DGAT1}}\right)\text{reference sample}}$$

However, it is difficult to find a standard sample as a reference which has copy number variation. Therefore, firstly, the reference was selected randomly. The copy number of reference was assumed as 1 copy, then calculate the accuracy of qPCR. After that, the copy number of reference was assumed as 2, and 3 and did the same process again. Because the

gene has more than 4 copies is rare, no more assumption was set up. By comparing the accuracy between the copy numbers 1, 2, 3 the copy number which has highest correction rate is considered as the correct copy number. Of course, this method could have bias since the assumption of copy number of reference could be wrong. The copy number evaluation thresholds table is given below (Table 6.1). Based on hypothetical copy number, using the copy number value calculated by above equation of each sample, the actual copy number of each sample can be evaluated.

Table 6.1 Hypothetical copy numbers of the reference and their thresholds (based on qPCR) for copy number evaluation.

Hypothetical copy number of the reference sample	1 copy	2 copies	3 copies	4 copies
1 copy	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5
2 copies	0.25-0.75	0.75-1.25	1.25-1.75	1.75-2.25
3 copies	0-0.459	0.459-0.825	0.825-1.165	1.165-1.495

6.3.5 Gene annotation

By using a custom-written python script, the position information of CNVs and SNPs was matched to Oar_v3.1 assembly, in order to get the names of genes (Ensemble database). Then all the gene names were input into online tools, bioDBnet (biodbnet-abcc.ncifcrf.gov/db/db2db.php), to get information of gene function. QTL information was obtained from online database, animal QTL db (https://www.animalgenome.org/cgi-bin/QTLdb/index) Ontology and Pathway analysis was done using PANTHER database (http://www.pantherdb.org)

6.3.6 Detection of selection signatures using SNP haplotypes

Selection signatures were detected by calculating the allele-specific extended haplotype homozygosity (EHH) within family line as well as the site-specific extended haplotype homozygosity (EHHS) between family lines, using a R package, REHH 2.0 (Gautier et al. 2017) (Appendix 6.4).

6.3.6.1 Within line allele-specific EHH test

6.3.6.1.1 Allele-specific EHH

Allele-specific EHH was employed to measure the extent to which an extended haplotype, involving a core allele (mutant/derived or ancestral), had been transmitted without any recombination (Sabeti et al. 2002) (Equation 6.1). The employed R package, REHH 2.0 (Gautier et al. 2017), estimated within-line EHH and the test statistic was iHS (Gautier and Naves 2011), the standardised ratio of the integrated allele-specific EHH (iHH).

Equation 6.1 Allele-specific EHH, where $K_{a_{s,t}}$ represents the number of distinct haplotypes (extending from SNP s to SNP t) carrying the core allele α_s, n_k is the observed count for the kth haplotype

$$\mathrm{EHH}_{\alpha_s,t} = \frac{1}{n_{\alpha_s}(n_{\alpha_s}-1)} \sum_{K=1}^{K_{a_{s,t}}} n_k\,(n_k - 1) \quad \text{(Gautier et al. 2017)}$$

$n_{a_s} = \sum_{K=1}^{K_{a_{s,t}}} n_k$, gives the total number of haplotypes carrying the core allele α_s.

An illustration of how EHH was estimated using the equation 6.1 is provided below (Gautier et al. 2017).

Normally the SNP marker is considered as biallelic, meaning there are only two kinds of base variants at each SNP locus (C→T/G→A) (Wang et al. 1998). Therefore, the variants at each

locus can be divided into two kinds, ancestral allele and derived allele. Consider eight copies (designated chromosome #1 to #8) of a segment of chromosome genotyped at 11 SNP loci, each spaced 1000 bp (1 kb) apart (Figure 6.1). At each locus, filled circle represents a derived (new) allele and bared line represents ancestral allele. EHH is calculated separately on ancestral allele and derived allele as EHH_{anc} and EHH_{der}, respectively. To do this, a focal SNP must be chosen. In this example, SNP located at 6 kb position is considered as focal SNP. Haplotype can be formed by extending this locus to left or right with different distance. Suppose haplotypes is formed by extending from the focal SNP (6 kb) to 7 kb. Based on the two kinds of variants at the focal locus, the chromosomes are divided into two lines, ancestral line (chromosome #1 to #4) and derived line (chromosome # 5 to #8). Therefore, EHH_{anc} and EHH_{der} is calculated separately for ancestral line and derived line. In the calculation of EHH_{anc} (chromosomes #1 to #4), there are two different haplotypes: bared line and bared line (chromosomes #1 and #4) and bared line and green circle (chromosomes #2 and #3). Therefore, $K_{a_{s,t}}=K_{anc,7}=2$. Besides, there are 2 chromosomes for each haplotype, so $n_1 = 2, n_2 = 2$. Finally, the calculation of $EHH_{anc,7}$ is shown below.

$$\sum_{K=1}^{K_{a_{s,t}}} n_k = \sum_{K=1}^{2} n_k = n_2 + n_1 = 2 + 2 = 4$$

$$EHH_{\alpha_s,t} = EHH_{anc,7} = \frac{1}{n_{anc}(n_{anc}-1)} \sum_{K=1}^{K_{anc,7}} n_k\,(n_k - 1)$$

$$= \frac{1}{\sum_{K=1}^{K_{a_{s,t}}} n_k \left(\sum_{K=1}^{K_{a_{s,t}}} n_k - 1\right)} \sum_{K=1}^{K_{anc,7}} n_k\,(n_k - 1)$$

$$= \frac{1}{4 \times 3}[2 \times (2-1) + 2 \times (2-1)] = \frac{1}{3}$$

In case of $EHH_{der,7}$ (chromosomes # 5 to #8), there is only one haplotype (bared line and bared line) involving the SNPs at 6 kb and 7 kb positions. Hence, $EHH_{der,7}$ estimated using the Equation 6.1 is equal to 1 (details shown below).

$$\sum_{K=1}^{K_{a_{s,t}}} n_k = \sum_{K=1}^{1} n_k = n_1 = 4$$

$$EHH_{\alpha_{s,t}} = EHH_{der,7} = \frac{1}{n_{der}(n_{der}-1)} \sum_{K=1}^{K_{der,7}} n_k\,(n_k - 1)$$

$$= \frac{1}{\sum_{K=1}^{K_{a_{s,t}}} n_k \left(\sum_{K=1}^{K_{a_{s,t}}} n_k - 1\right)} \sum_{K=1}^{K_{der,7}} n_k\,(n_k - 1)$$

$$= \frac{1}{4 \times 3}[4 \times (4-1)] = 1$$

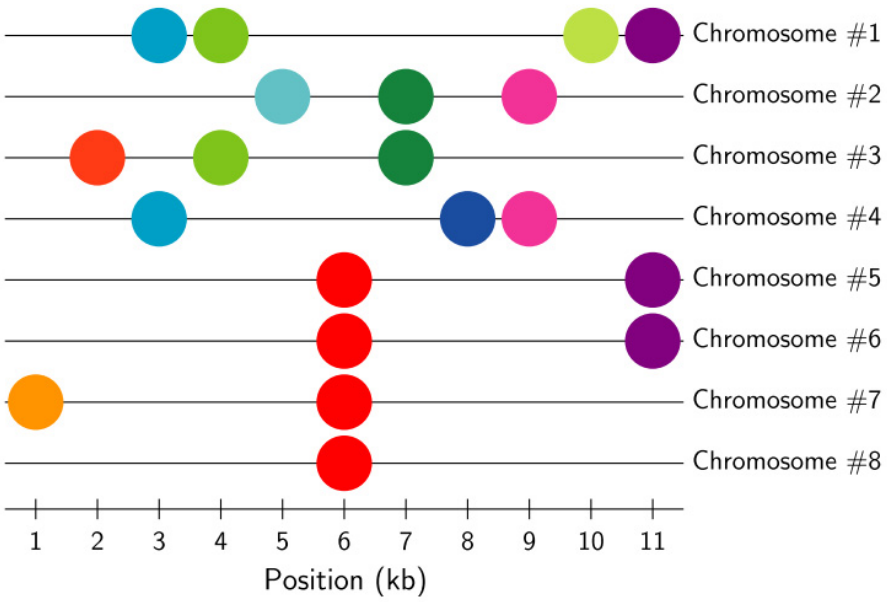


Figure 6.1 Schematic view of 11 SNPs in eight aligned chromosomes.

Each of the eight lines symbolises a chromosome and a filled circle represents a derived allele at the corresponding SNP position (Gautier et al. 2017)

6.3.6.1.2 Integrated EHH (iHH)

The next step after estimating EHH_{anc} and EHH_{der} was estimation of integrated EHH (iHH), Similar to EHH, there are two kinds of iHH, iHHa and iHHd (corresponding to the ancestral and derived alleles, respectively), which are defined as the area under the EHH_{anc} and EHH_{der} curves, respectively (illustrated in Figure 6.2).

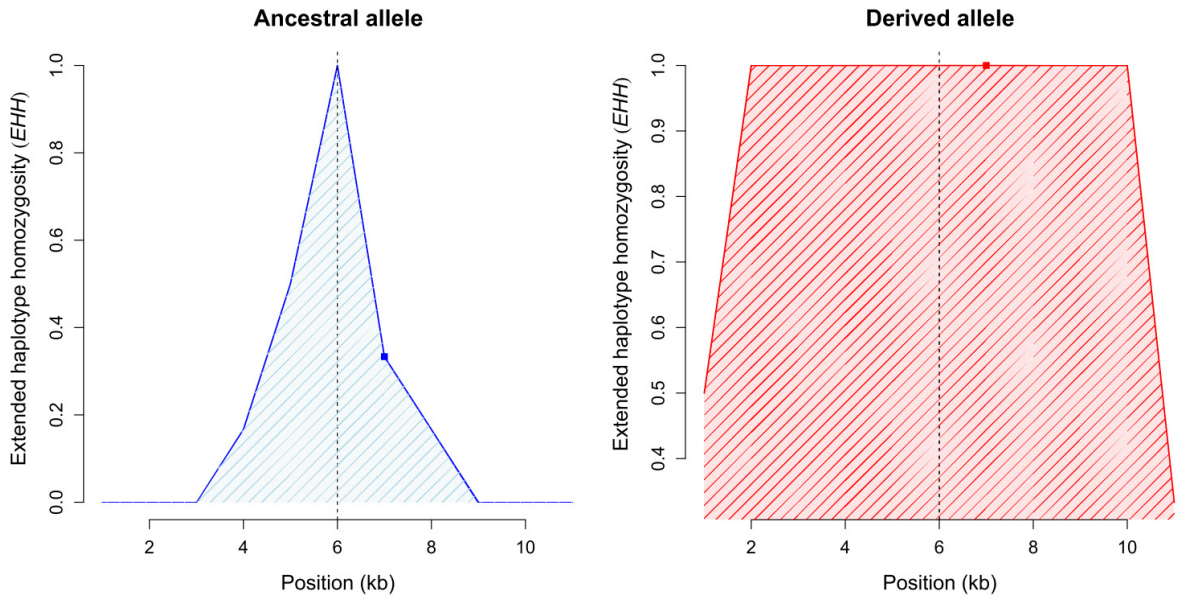


Figure 6.2 Illustration of iHH (shaded part) (Gautier et al. 2017).

The squared dots represent the example values of EHH computed above. In each plot, the shaded area represents the integrated EHH (iHH) (plot to the left is iHHa and to the right is iHHd). The EHH decays far more rapidly for the haplotypes carrying the ancestral variant at the core SNP (blue curve on the left-hand side) than for those carrying the derived variant (red curve, on the right-hand side).

6.3.6.1.3 UniHS

Subsequently, UniHS, log ratio of the iHH for its ancestral (iHH_a) and derived (iHH_d) alleles (Voight et al. 2006), was calculated (Equation 6.2).

Equation 6.2 UniHS calculation

$$\mathrm{UniHS} = \mathrm{Log}\left(\frac{\mathrm{iHH_a}}{\mathrm{iHH_d}}\right)$$

6.3.6.1.4 Standardised ratio of iHH (iHS)

Finally, the test-statistic, standardised ratio of iHH (iHS) for a given focal SNP was computed (Voight et al. 2006) as shown below (Equation 6.3).

Equation 6.3 iHS calculation, where $\mu_{\mathrm{UniHS}}^{\mathrm{p_s}}$ and $\sigma_{\mathrm{UniHS}}^{\mathrm{p_s}}$represent, respectively, the average and the standard deviation of the UniHS computed over all the SNPs with a derived allele frequency $\mathrm{p_s}$ similar to that of the core SNPs.

$$\mathrm{iHS(s)} = \frac{\mathrm{UniHS(s)} - \mu_{\mathrm{UniHS}}^{p_s}}{\sigma_{\mathrm{UniHS}}^{p_s}}$$

6.3.6.1.5 P*iHS*

*P*iHS, a two-sided p-value (in a -log10 scale) associated with the null hypothesis of selective neutrality, was estimated as shown below (Equation 6.4).

Equation 6.4 *P*iHS calculation, where $\emptyset(x)$ represents the Gaussian cumulative distribution function.

$$p_{iHS=-log_{10}(1-2|\emptyset(iHS)-0.5|)}$$

6.3.6.2 Between-line site-specific extended haplotype homozygosity (EHHS)

Within-line allele-specific EHH test might have low power of detecting positive selection signatures at a high frequency of the selected allele in population, especially when fixed, and a between-population site-specific EHH (EHHS) approach was proposed to overcome such limitation (Tang et al. 2007). EHHS method would compare the decay of EHH of an individual SNP site, instead of EHH of an allele, across populations or lines. Two separate test statistics were independently proposed for EHHS: xp-EHH (Sabeti et al. 2007) and Rsb (Tang et al. 2007). In the current study, between-line xp-EHH and Rsb were estimated using REHH 2.0 (Gautier et al. 2017) and the underlying computations of the statistics is illustrated below.

The calculation of $\mathrm{EHHS}_{s,t}^{Sabeti}$ (Equation 6.5) and $\mathrm{EHHS}_{s,t}^{Tang}$ is shown below.

Equation 6.5 Sabeti's EHHS calculation (Sabeti et al. 2007), where $K_{a_s,t}$ n_s and n_k represent the number of distinct alleles, the total number of haplotypes carrying the core allele a_s, and the observed count for the kth haplotype, respectively.

$$\mathrm{EHHS}_{s,t}^{Sabeti} = \frac{1}{n_s(n_s-1)} \sum_{a_s=1}^{2} \left(\sum_{k=1}^{K_{a_s,t}} n_k(n_k-1) \right)$$

$$= \frac{1}{n_s(n_s-1)} \left[\sum_{k=1}^{K_{anc,t}} n_k(n_k-1) + \sum_{k=1}^{K_{der,t}} n_k(n_k-1) \right]$$

$$n_s = \sum_{a_s=1}^{2} n_{a_s}$$

Equation 6.6 Tang's EHHS calculation (Tang et al. 2007)

$$\mathrm{EHHS}_{s,t}^{Tang} = \frac{1-h_{hap}^{(s,t)}}{1-h_{all}^{(s)}}$$

$$h_{all}^{(s)} = \frac{n_s}{n_s-1} \left(1 - \frac{1}{n_s^2} \sum_{a_s=1}^{2} n_{a_s}^2 \right)$$

$$h_{hap}^{(s,t)} = \frac{n_s}{n_s-1} \left(1 - \frac{1}{n_s^2} \sum_{a_s=1}^{2} \left(\sum_{k=1}^{K_{a_s,t}} n_k^2 \right) \right)$$

$h_{all}^{(s)}$ is an estimator of the focal SNP heterozygosity and $h_{hap}^{(s,t)}$ is an estimator of haplotype heterozygosity over the chromosome interval extending from SNPs to SNP. Besides, $K_{a_s,t}$, n_s and n_k represent the number of distinct alleles, the total number of haplotypes carrying the core allele a_s, and the observed count for the kth haplotype, respectively.

6.3.6.2.1 iES (The integrated EHHS)

iES is defined as the area under the EHHS curve with respect to map position as Figure 6.3.

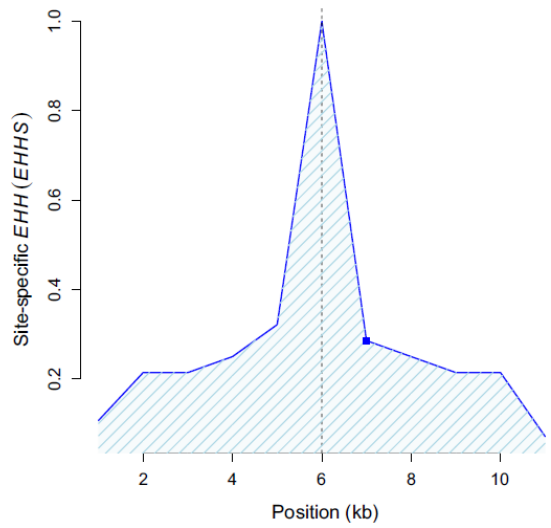


Figure 6.3 illustrate of iES (shadow part) (Gautier et al. 2017).

The squared dots represent the example values of EHHS computed above. In each Figure, the shaded area represents the integrated EHHS (iHH).

6.3.6.2.2 XP-EHH and Rsb (The standardized ratios of pairwise line iES)

$\mathrm{LRiES}^{\mathrm{Sabeti}}(\mathrm{s})$ and $\mathrm{LRiES}^{\mathrm{Tang}}(\mathrm{s})$ were calculated using Equation 6.7 and Equation 6.8, then XP-EHH and Rsb using Equation 6.9 and Equation 6.10.

Equation 6.7 Sabeti's LRiES calculation, where $\mathrm{iES}_{\mathrm{pop1}}^{\mathrm{Sabeti}}(\mathrm{s})$ is the iES of population 1 calculated by Sabeti while $\mathrm{iES}_{\mathrm{pop2}}^{\mathrm{Sabeti}}(\mathrm{s})$ is the iES of population 2

$$\mathrm{LRiES}^{\mathrm{Sabeti}}(\mathrm{s}) = \log\left(\frac{\mathrm{iES}_{\mathrm{pop1}}^{\mathrm{Sabeti}}(\mathrm{s})}{\mathrm{iES}_{\mathrm{pop2}}^{\mathrm{Sabeti}}(\mathrm{s})}\right)$$

Equation 6.8 Tang's LRiES calculation, where $\mathrm{iES}_{\mathrm{pop1}}^{\mathrm{Tang}}(\mathrm{s})$ is the iES of population 1 calculated by Tang while $\mathrm{iES}_{\mathrm{pop2}}^{\mathrm{Tang}}(\mathrm{s})$ is the iES of population 2.

$$\mathrm{LRiES}^{\mathrm{Tang}}(\mathrm{s}) = \log\left(\frac{\mathrm{iES}_{\mathrm{pop1}}^{\mathrm{Tang}}(\mathrm{s})}{\mathrm{iES}_{\mathrm{pop2}}^{\mathrm{Tang}}(\mathrm{s})}\right)$$

Equation 6.9 XP-EHH and p(xp-EHH) calculation, where $\emptyset(\mathrm{XP}-\mathrm{EHH})$ represents the Gaussian cumulative distribution function, $\mathrm{med}_{\mathrm{LRiES^{Sabeti}}}$ and $\sigma_{\mathrm{LRiES^{Sabeti}}}$ the average and the standard deviation of the $\mathrm{LRiES^{Sabeti}}$

$$\mathrm{XP}-\mathrm{EHH(s)} = \frac{\mathrm{LRiES^{Sabeti}(s)} - \mathrm{med}_{\mathrm{LRiES^{Sabeti}}}}{\sigma_{\mathrm{LRiES^{Sabeti}}}}$$

$$p_{\mathrm{XP-EHH}} = -\log_{10}\left(1 - 2\left|\emptyset_{(\mathrm{XP-EHH})} - 0.5\right|\right)$$

Equation 6.10 Rsb and p(Rsb) calculation, where $\emptyset(\mathrm{Rsb})$ represents the Gaussian cumulative distribution function, $\mathrm{LRiES^{Tang}}$ and $\sigma_{\mathrm{LRiES^{Tang}}}$ the average and the standard deviation of the $\mathrm{LRiES^{Tang}}$

$$\mathrm{Rsb(s)} = \frac{\mathrm{LRiES^{Tang}(s)} - \mathrm{med}_{\mathrm{LRiES^{Tang}}}}{\sigma_{\mathrm{LRiES^{Tang}}}}$$

$$p_{\mathrm{Rsb}} = -\log_{10}(1 - 2|\emptyset_{(\mathrm{Rsb})} - 0.5|)$$

6.3.6.2.3 Selection signature region

Selective signature regions (Gautier et al. 2017) were built by customized script using python (Appendix 6.2). The chromosome 13 was divided into 333 consecutive 500 kb windows (with a 250kb overlap). Windows with at least 2 SNPs displaying an absolute value of the -log10 (p-value) > 4 for at least one of the four test statistics were considered significant.

6.4 Results

6.4.1 Quality control

Of the total 606,005 SNPs located on the 26 autosomes, 463,392 SNPs passed the quality control threshold. In total, all 93 Romney (42 resistant and 51 resilient) sheep passed the quality control (Additional file: Table S6.1 sample information).

6.4.2 CNVs and CNVRs

In total, 3313 CNVs (1833 losses and 1480 gains) were found. The size of CNVs ranged from 64 bp to 350 kb, with a median of 15.2 kb and an average of 30.5 kb. By merging CNVs of

all samples, totally 399 CNVRs were created. Of them, 157 were gains, 200 losses and 42 mixes. The size of CNVRs ranged from 64 bp to 198.2 kb, with a median of 10.1 kb and an average of 20.8 kb.

In the resilient line, 1805 CNVs were detected and from these, 314 CNVRs (137 gains, 145 losses and 32 mixes) were formed. The observed CNV and CNVR numbers in the resistant line were 1,508 and 293 (112 gains, 153 losses and 28 mixes), respectively. Chromosome-wise distribution of CNVRs in the two lines is depicted in Figure 6.4. Only 196 CNVRs were common between the two lines (Figure 6.5).

6.4.3 SNP-based selection signatures

6.4.3.1 Within-line allele-specific extended haplotype homozygosity (EHH)

The EHH test was done within the resistant and resilient lines, to detect positive selection signatures. Three threshold levels ($P<0.01$, $P<0.001$ and $P<0.0001$) of *P*iHS were considered. The number of SNPs found to be significant at respective threshold levels were 2,934, 379 and 85 in the resilient line and 2,922, 347 and 62 in the resistant line. A plot, based on iHS on chr2, showing the difference between two lines is shown in Figure 6.6, while those for other chromosomes were included in Appendix 6.3. Chromosome-wise number of significant SNPs at the three threshold levels are presented in Table 6.2. Exact location details, iHS and PiHS pertaining to the SNPs significant at P<0.0001 level, in the resilient and resistant lines, are presented in Table 6.3 and Table 6.4, respectively; those pertaining to SNPs significant at $P<0.01$ and $P<0.001$ levels are presented in additional files (Tables S6.2 and S6.3). At the hardest threshold level, $P < 0.0001$, there was no SNP shared between two lines.

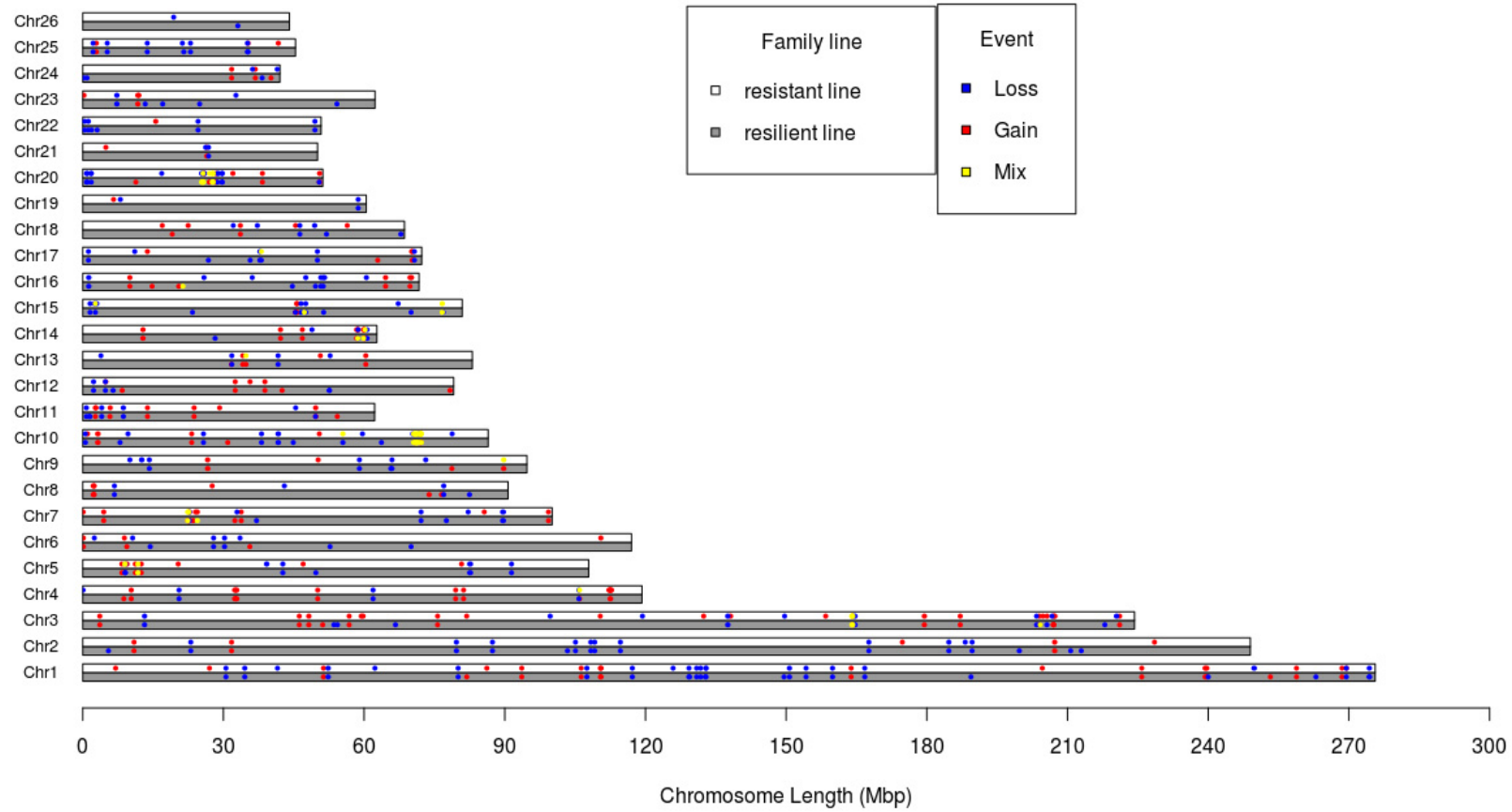


Figure 6.4 Chromosomal distribution of copy number variant regions (CNVR) detected in gastrointestinal nematode resistant (white bars) and resilient (grey bars) lines of Romney sheep.

CNVR losses, gains and mix were depicted in blue, red and yellow, respectively.

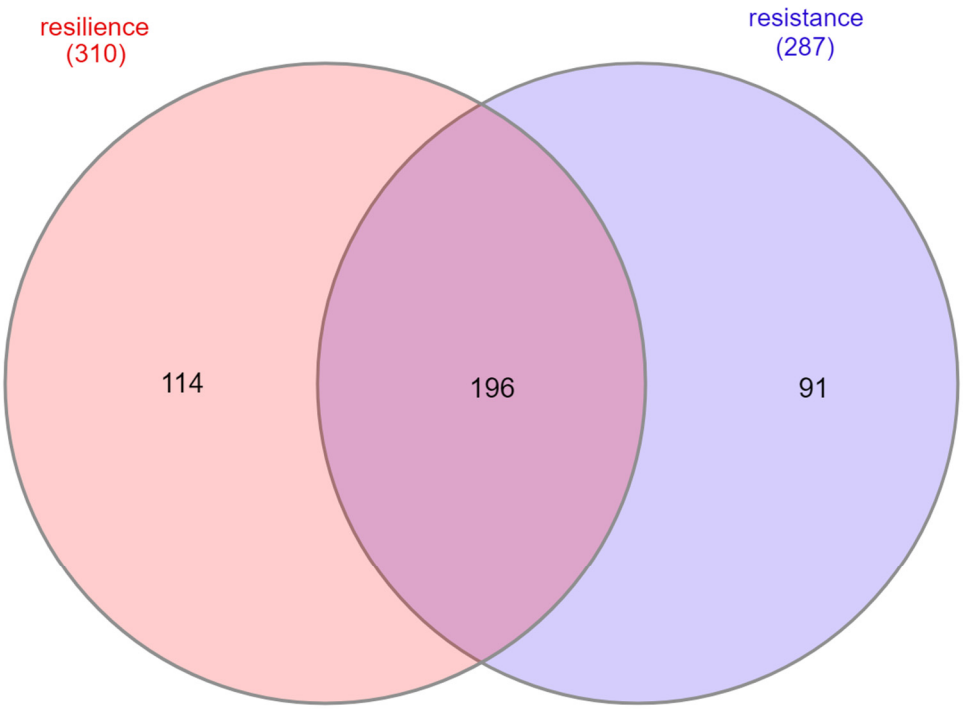


Figure 6.5 Venn plot of copy number variant regions (CNVR) detected in gastrointestinal nematode resistant (green circle) and resilient (orange) lines of Romney sheep.

Number in overlapping regions denote the number of CNVR common to both lines while those in non-overlapping regions are unique for each line.

6.4.3.2 Between-line site-specific extended haplotype homozygosity (EHHS)

The EHHS test was done between the resistant and resilient lines, to detect positive selection signatures. Three threshold levels (P<0.01, P<0.001 and P<0.0001) of two algorithms, XP-EHH and Rsb, were considered. The number of SNPs found to be significant at respective threshold levels were 2,947, 236 and 39 using XP-EHH and 4064, 322 and 48 using Rsb (Table 6.6).

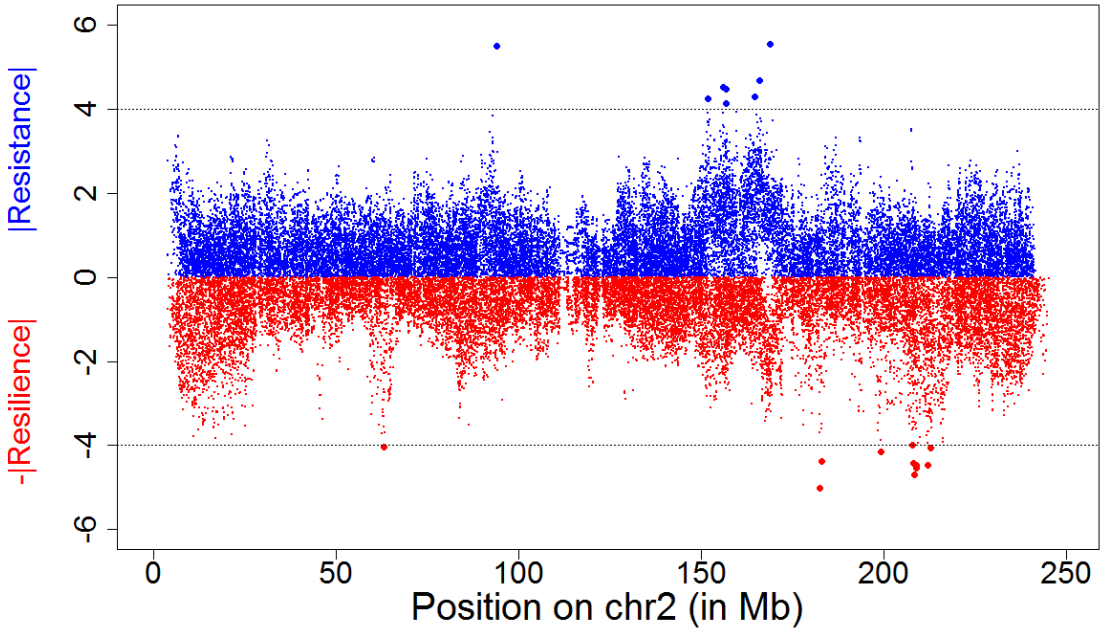


Figure 6.6 Plot showing the differences in iHS, the within-line allele-specific extended haplotype homozygosity (EHH) test statistic, with regard to single nucleotide polymorphism (SNP) loci located on chromosome 2 between two Romney sheep lines (gastrointestinal nematode resistant and resilient).

The blue and red points represent SNPs of in resistant and resilient lines, respectively. The absolute value on Y axis represents log(p-value). The two grey lines represent log(p-value) thresholds for each line and the bold points represent the SNPs that passed the threshold (dotted line). X axis represents position on the chromosome.

The SNPs detected using Rsb and XP-EHH at threshold p <0.0001 are shown in Table 6.7 and Table 6.8 (full results were included in additional files, Table S6.4 and Table S6.5), respectively. Of those, only 10 SNPs were shared between the two algorithms (Figure 6.7) and those were located on chr11 and 13 (Table 6.9).

6.4.3.3 Selection signature regions

Two selection signature regions were found on chromosome 13 (Table 6.10). There were 9 genes locating in those regions.

6.4.4 qPCR validation

Four randomly selected CNV segments were validated by qPCR, using DNA from 8 samples. Validation accuracy with regard to the four markers ranged between 50.0% to 100.0%, with an overall mean of 73.33% (Table 6.5; Additional file: S6.8)

Table 6.2 Results of within-line allele-specific EHH test in gastrointestinal nematode resilient and resistant lines of Romney sheep: chromosome-wise number of SNPs evincing signatures of selection

Chr	Resilience			Resistance		
	p< 0.01	p< 0.001	p<0.0001	p< 0.01	p< 0.001	p<0.0001
1	262	27	3	431	45	5
2	612	102	17	449	60	12
3	280	15	4	282	19	3
4	161	12	2	159	17	1
5	125	14	5	123	21	4
6	106	14	3	108	12	1
7	139	18	1	85	11	2
8	100	6	1	122	20	10
9	124	12	1	117	18	2
10	82	10	1	87	12	7
11	65	9	2	67	7	0
12	113	15	3	81	3	1
13	128	16	2	65	5	2
14	80	17	6	59	9	1
15	88	17	7	105	15	1
16	45	8	3	52	0	0
17	124	21	1	70	19	5
18	54	11	3	66	3	0
19	31	0	0	68	4	0
20	23	2	0	51	6	1
21	39	12	10	62	15	2
22	27	4	2	61	3	0
23	40	5	4	64	17	2
24	29	8	4	29	0	0
25	23	1	0	34	4	0
26	34	3	0	25	2	0
Total	2934	379	85	2922	347	62

Table 6.3 List of SNPs detected by iHS and *P*iHS, found to significant (P<0.0001) positive selection signatures in the resilient line.

SNP	CHR	Position	Gene	QTL ID	QTL Description
oar3_OAR1_46230979	1	46230979			
oar3_OAR1_46251958	1	46251958			
oar3_OAR1_92287324	1	92287324			
oar3_OAR2_63269803	2	63269803	ALDH1A1		
s43294.1	2	182500225			
oar3_OAR2_183118980	2	183118980			
oar3_OAR2_199403260	2	199403260			
oar3_OAR2_199403739	2	199403739			
oar3_OAR2_207802788	2	207802788			
oar3_OAR2_207816125	2	207816125			
oar3_OAR2_208220289	2	208220289			
oar3_OAR2_208379447	2	208379447			
oar3_OAR2_209031436	2	209031436			
oar3_OAR2_209063396	2	209063396			
oar3_OAR2_209380137	2	209380137			
oar3_OAR2_212034902	2	212034902			
oar3_OAR2_212143325	2	212143325	ERBB4		
oar3_OAR2_212971849	2	212971849			
oar3_OAR2_216268337	2	216268337			
oar3_OAR2_216275404	2	216275404	ATIC		
oar3_OAR3_30424816	3	30424816			
s60811.1	3	112141245			
oar3_OAR3_112153296	3	112153296			
oar3_OAR3_168208985	3	168208985	ENSOART00000014456		
oar3_OAR4_60004430	4	60004430	ELMO1		
oar3_OAR4_111199347	4	111199347		QTL:19803	Haemonchus contortus FEC
oar3_OAR5_20899771	5	20899771			
oar3_OAR5_20923573	5	20923573			
oar3_OAR5_81415565	5	81415565	EDIL3		
oar3_OAR5_81867785	5	81867785			
oar3_OAR5_85029774	5	85029774			

SNP	CHR	Position	Gene	QTL ID	QTL Description
oar3_OAR6_16496167	6	16496167	COL25A1		
oar3_OAR6_52244068	6	52244068		QTL:16024	Fecal egg count
oar3_OAR6_52629465	6	52629465		QTL:16024	Fecal egg count
oar3_OAR7_13658773	7	13658773		QTL:95614	Fecal egg count
oar3_OAR8_46694405	8	46694405		QTL:12899	Trichostrongylus adult and larva
oar3_OAR9_81364803	9	81364803			
oar3_OAR10_74818035	10	74818035	STK24	QTL:13989	Fecal egg count
oar3_OAR11_23624585	11	23624585		QTL:12901	Trichostrongylus adult and larva
oar3_OAR11_23841741	11	23841741	ASPA	QTL:12901	Trichostrongylus adult and larva
oar3_OAR12_15910699	12	15910699			
oar3_OAR12_18022014	12	18022014			
oar3_OAR12_20839567	12	20839567			
oar3_OAR13_71532542	13	71532542		QTL:16027	Fecal egg count
oar3_OAR13_72500964	13	72500964		QTL:16027	Fecal egg count
s11567.1	14	8824045	CDH13	QTL:12893	Nematodirus FEC
oar3_OAR14_24936118	14	24936118	CCDC102A	QTL:16028	Fecal egg count
oar3_OAR14_42469160	14	42469160	NUDT19	QTL:12893	Nematodirus FEC
oar3_OAR14_42471533	14	42471533		QTL:12893	Nematodirus FEC
oar3_OAR14_42723029	14	42723029	RHPN2	QTL:12893	Nematodirus FEC
oar3_OAR14_42753314	14	42753314	RHPN2	QTL:12893	Nematodirus FEC
oar3_OAR15_7478160	15	7478160	ARHGAP42		
oar3_OAR15_8612651	15	8612651	CNTN5		
oar3_OAR15_15035566	15	15035566			
oar3_OAR15_28418100	15	28418100	MPZL3	QTL:16029	Fecal egg count
oar3_OAR15_28424239	15	28424239	MPZL3	QTL:16029	Fecal egg count
oar3_OAR15_28443572	15	28443572	MPZL2	QTL:16029	Fecal egg count
oar3_OAR15_76299782	15	76299782	ENSOART00000009126		
oar3_OAR16_5773509	16	5773509			
oar3_OAR16_20999184	16	20999184			
oar3_OAR16_21000031	16	21000031			
oar3_OAR17_15472520	17	15472520	INPP4B	QTL:16031	Fecal egg count
oar3_OAR18_42487138	18	42487138			
oar3_OAR18_42577999	18	42577999			
oar3_OAR18_42585260	18	42585260			

SNP	CHR	Position	Gene	QTL ID	QTL Description
oar3_OAR21_23365859	21	23365859	ENSOART00000008626		
oar3_OAR21_23385917	21	23385917	ENSOART00000008626		
oar3_OAR21_23419044	21	23419044	ENSOART00000008626		
oar3_OAR21_23459252	21	23459252	ENSOART00000008626		
oar3_OAR21_23460758	21	23460758	ENSOART00000008626		
oar3_OAR21_24609991	21	24609991			
oar3_OAR21_24732343	21	24732343			
oar3_OAR21_24736972	21	24736972			
oar3_OAR21_24877528	21	24877528			
oar3_OAR21_27649204	21	27649204	CCDC15		
oar3_OAR22_41630273	22	41630273	ENSOART00000009611		
oar3_OAR22_41940069	22	41940069			
oar3_OAR23_36595784	23	36595784			
oar3_OAR23_49810104	23	49810104			
oar3_OAR23_50091419	23	50091419			
oar3_OAR23_52340197	23	52340197			
oar3_OAR24_24300131	24	24300131			
oar3_OAR24_24402575	24	24402575			
oar3_OAR24_24412081	24	24412081	ENSOART00000027850		
oar3_OAR24_24459152	24	24459152			

Table 6.4 List of SNPs detected by iHS and PiHS, found to evince significant (P<0.0001) positive selection signatures in the resistant line.

SNP	CHR	Position	Gene	QTL ID	QTL Description
oar3_OAR1_78406214	1	78406214			
oar3_OAR1_78418806	1	78418806			
oar3_OAR1_109434615	1	109434615			
oar3_OAR1_109497543	1	109497543	ATP1A2		
oar3_OAR1_171568286	1	171568286	HHLA2		
oar3_OAR2_94007703	2	94007703			
OAR2_160946331.1	2	151793561			
oar3_OAR2_151911086	2	151911086	GPD2		
oar3_OAR2_155966958	2	155966958	FMNL2		
oar3_OAR2_156131938	2	156131938			
oar3_OAR2_156132069	2	156132069			
oar3_OAR2_156978779	2	156978779			
oar3_OAR2_156997706	2	156997706			
oar3_OAR2_159759978	2	159759978			
oar3_OAR2_164695219	2	164695219	GTDC1		
oar3_OAR2_165998566	2	165998566	KYNU		
oar3_OAR2_168813588	2	168813588	LRP1B		
OAR3_79055518.1	3	74855304			
oar3_OAR3_74858149	3	74858149			
oar3_OAR3_190853439	3	190853439			
OAR4_19365053.1	4	18993547		QTL:19803	Haemonchus
oar3_OAR5_21091255	5	21091255			
oar3_OAR5_83171184	5	83171184			
oar3_OAR5_83172657	5	83172657			
oar3_OAR5_83777203	5	83777203			
oar3_OAR6_10465659	6	10465659			
oar3_OAR7_26775738	7	26775738			
oar3_OAR7_26810566	7	26810566			
OAR8_15071235.1	8	13501674			
oar3_OAR8_16444613	8	16444613			
OAR8_18319939.1	8	16445482			
oar3_OAR8_16461549	8	16461549			

SNP	CHR	Position	Gene	QTL ID	QTL Description
oar3_OAR8_16756030	8	16756030			
oar3_OAR8_17168436	8	17168436			
oar3_OAR8_17175021	8	17175021			
oar3_OAR8_51154733	8	51154733		QTL:12899	Trichostrongylus
oar3_OAR8_51155071	8	51155071		QTL:12899	Trichostrongylus
oar3_OAR8_82213039	8	82213039	FNDC1	QTL:16025	Fecal egg count
OAR9_5518009.1	9	5641037	ADGRB3		
oar3_OAR9_8364653	9	8364653			
oar3_OAR10_52190090	10	52190090		QTL:13989	Fecal egg count
oar3_OAR10_55315348	10	55315348		QTL:13989	Fecal egg count
oar3_OAR10_78855201	10	78855201		QTL:13989	Fecal egg count
oar3_OAR10_79112972	10	79112972		QTL:13989	Fecal egg count
oar3_OAR10_79614660	10	79614660		QTL:13989	Fecal egg count
oar3_OAR10_79774280	10	79774280	ENSOARG00000005401	QTL:13989	Fecal egg count
oar3_OAR10_79991716	10	79991716		QTL:13989	Fecal egg count
oar3_OAR12_36748585	12	36748585	MROH9		
oar3_OAR13_4478216	13	4478216			
OAR13_5326638.1	13	4480720			
oar3_OAR14_47857530	14	47857530		QTL:12893	Nematodirus FEC
oar3_OAR15_57891899	15	57891899			
oar3_OAR17_36538740	17	36538740	FSTL5		
oar3_OAR17_36559268	17	36559268	FSTL5		
oar3_OAR17_42455943	17	42455943			
oar3_OAR17_42604405	17	42604405			
oar3_OAR17_42612721	17	42612721			
oar3_OAR20_43569986	20	43569986			
oar3_OAR21_8407961	21	8407961	ME3		
oar3_OAR21_25979208	21	25979208			
oar3_OAR23_10112869	23	10112869		QTL:19791	Haemonchus
oar3_OAR23_50253530	23	50253530			

Table 6.5 Results of qPCR validation of 4 CNVs

Primer ID	Samples validated	Validation accuracy	
		Proportion	Percentage
J26	6	6/6	100.0%
J27	8	4/8	50.0%
J28	8	5/8	62.5%
J29	8	7/8	87.5%
Total	30	22/30	73.33%

Table 6.6 Results of between-line EHHS test (using two different algorithms, XP-EHH and Rsb) in gastrointestinal nematode resilient and resistant lines of Romney sheep: chromosome-wise number of SNPs evincing signatures of selection.

Chr	Rsb			XP-EHH		
	p< 0.01	p< 0.001	p<0.0001	p< 0.01	p< 0.001	p<0.0001
1	424	14	0	57	0	0
2	434	62	4	599	34	0
3	269	4	0	59	0	0
4	219	21	4	169	0	0
5	174	3	0	77	1	0
6	194	4	0	5	0	0
7	158	5	0	73	0	0
8	162	0	0	28	1	0
9	89	0	0	15	0	0
10	196	20	0	729	12	0
11	110	23	5	250	58	7
12	72	1	0	0	0	0
13	262	61	19	227	117	32
14	30	1	0	0	0	0
15	160	5	0	0	0	0
16	103	12	0	181	1	0
17	162	11	0	59	0	0
18	162	31	12	41	12	0
19	100	3	0	35	0	0
20	133	12	0	77	0	0
21	77	4	0	39	0	0
22	94	3	0	34	0	0
23	72	2	0	0	0	0
24	73	14	4	168	0	0
25	63	0	0	0	0	0
26	72	6	0	25	0	0
Total	4064	322	48	2947	236	39

Table 6.7 Significant (p <0.0001) SNPs detected using Rsb.

SNP_ID	Chr	Position	Gene	Gene function	QTL ID	QTL function
oar3_OAR2_71491461	2	71491461	None		QTL:12898	Trichostrongylus adult and larva count
oar3_OAR2_71709132	2	71709132	GLIS3	GLIS family zinc finger 3	QTL:12898	richostrongylus adult and larva coun
oar3_OAR2_156978779	2	156978779	None			
oar3_OAR2_157065505	2	157065505	LOC101112890			
oar3_OAR4_111118668	4	111118668	None			
oar3_OAR4_111119582	4	111119582	None			
oar3_OAR4_111125648	4	111125648	None			
oar3_OAR4_111167177	4	111167177	None			
oar3_OAR11_47837547	11	47837547	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_47864207	11	47864207	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_47889283	11	47889283	PECAM1	platelet/endothelial cell adhesion molecule 1	QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_47920657	11	47920657	PECAM1	platelet/endothelial cell adhesion molecule 1	QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48327544	11	48327544	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR13_70757305	13	70757305	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70758173	13	70758173	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70810243	13	70810243	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70820259	13	70820259	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70853062	13	70853062	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count

SNP_ID	Chr	Position	Gene	Gene function	QTL ID	QTL function
oar3_OAR13_70853714	13	70853714	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70870621	13	70870621	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70876794	13	70876794	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70887333	13	70887333	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70891326	13	70891326	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70896117	13	70896117	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70927833	13	70927833	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_70930065	13	70930065	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_71026943	13	71026943	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_71028838	13	71028838	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count

SNP_ID	Chr	Position	Gene	Gene function	QTL ID	QTL function
oar3_OAR13_71249361	13	71249361	PTPRT	protein tyrosine phosphatase, receptor type, t	QTL:16027	Fecal egg count
oar3_OAR13_71405241	13	71405241	None		QTL:16027	Fecal egg count
oar3_OAR13_71450224	13	71450224	None		QTL:16027	Fecal egg count
oar3_OAR13_71615455	13	71615455	None		QTL:16027	Fecal egg count
oar3_OAR18_16432552	18	16432552	AGBL1	ATP/GTP binding protein-like 1		
oar3_OAR18_16462493	18	16462493	AGBL1	ATP/GTP binding protein-like 1		
oar3_OAR18_16492855	18	16492855	AGBL2	ATP/GTP binding protein-like 2		
oar3_OAR18_16617744	18	16617744	AGBL3	ATP/GTP binding protein-like 3		
oar3_OAR18_16899228	18	16899228	None			
oar3_OAR18_16926616	18	16926616	None			
oar3_OAR18_16960784	18	16960784	None			
oar3_OAR18_16963447	18	16963447	None			
s46270.1	18	16975561	None			
oar3_OAR18_16979598	18	16979598	None			
oar3_OAR18_56441194	18	56441194	GHGA	autocrine or paracrine negative modulators of the neuroendocrine system	QTL:12965	Haemonchus contortus FEC
oar3_OAR18_57318186	18	57318186	None		QTL:12965	Haemonchus contortus FEC
oar3_OAR24_24325647	24	24325647	None			
oar3_OAR24_24415558	24	24415558	None			
oar3_OAR24_24429370	24	24429370	None			
oar3_OAR24_24441245	24	24441245	None			

Table 6.8 Significant (p< 0.0001) SNPs detected by XPEHH.

SNP_ID	Chr	Position	Gene	Gene function	QTL ID	QTL function
oar3_OAR11_48266583	11	48266583	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48269342	11	48269342	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48303536	11	48303536	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48304970	11	48304970	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48327544	11	48327544	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48330513	11	48330513	None		QTL:12901	Trichostrongylus adult and larva count
oar3_OAR11_48340935	11	48340935	BPTF	bromodomain PHD finger transcription factor	QTL:12901	Trichostrongylus adult and larva count
oar3_OAR13_70728930	13	70728930	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70771651	13	70771651	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70772623	13	70772623	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70779221	13	70779221	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70788682	13	70788682	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70792222	13	70792222	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70792621	13	70792621	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70801328	13	70801328	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
s34281.1	13	70802486	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count

SNP_ID	Chr	Position	Gene	Gene function	QTL ID	QTL function
oar3_OAR13_70809606	13	70809606	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70810243	13	70810243	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70820259	13	70820259	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70823669	13	70823669	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70826211	13	70826211	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70837977	13	70837977	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70853062	13	70853062	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70853714	13	70853714	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70854996	13	70854996	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70856641	13	70856641	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70864938	13	70864938	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70870621	13	70870621	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70873340	13	70873340	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70876794	13	70876794	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70877534	13	70877534	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
s18276.1	13	70886211	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70887333	13	70887333	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count

SNP_ID	Chr	Position	Gene	Gene function	QTL ID	QTL function
oar3_OAR13_70890125	13	70890125	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70891326	13	70891326	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70896117	13	70896117	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70896623	13	70896623	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70906686	13	70906686	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count
oar3_OAR13_70906746	13	70906746	PTPRT	protein tyrosine phosphatase receptor type t	QTL:16027	Fecal egg count

Table 6.9 List of SNPs detected by both between-line EHHS algorithms, XP-EHH and Rsb, found to evince significant (P<0.0001) positive selection signatures

SNP	Chr	Position	Gene	QTL name	QTL ID	Description
oar3_OAR11_48327544	11	48327544	none	LATRICH_2	QTL:12901	larva count
oar3_OAR13_70810243	13	70810243	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70820259	13	70820259	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70853062	13	70853062	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70853714	13	70853714	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70870621	13	70870621	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70876794	13	70876794	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70887333	13	70887333	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70891326	13	70891326	PTPRT	FECGEN	QTL:16027	Fecal egg count
oar3_OAR13_70896117	13	70896117	PTPRT	FECGEN	QTL:16027	Fecal egg count

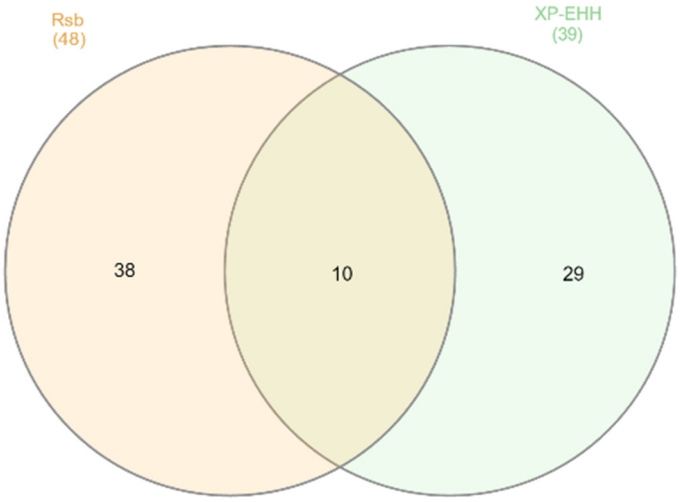


Figure 6.7 the difference of positive selection signature detected between XP-EHH & Rsb.

Pink and green circles represent positive selection signature (SNPs) by Rsb and XP-EHH respectively. Numbers in overlapping regions denote the number of common positive selection signature to both methods while those in non-overlapping regions are unique for each method.

Table 6.10 Selection signature regions on chromosome 13.

The chromosome 13 was divided into 333 consecutive 500 kb windows (with a 250 kb overlap). Windows with at least 2 SNPs displaying an absolute value of the -log10 (p-value) > 4 for at least one of the four test statistics were considered significant.

Region	Chr	Start	End	Gene	Test	Peak position	Peak value	SNP
1	13	4250000	4500000	None	iHS_resistance	4197047	-0.278628	2
					iHS_resilience	4008310	0	0
					rsb	4161323	0.2844913	0
					xpehh	4185377	0.4031644	0
		4500000	4750000		iHS_resistance	4647809	-0.652868	2
					iHS_resilience	4268677	0	0
					rsb	4614703	0.119337	0
					xpehh	4266104	-0.201568	0
2	13	70750000	71000000	PTPRT LOC106991533 SRSF6 L3MBTL1 SGK2 IFT52 LOC105610217 MYBL2 GTSF1L	iHS_resistance	70853062	-3.037707	0
					iHS_resilience	70757305	-2.180322	0
					rsb	70758173	4.9297797	13
					xpehh	70873340	4.8860907	32
		71000000	71250000		iHS_resistance	70853062	-3.037707	0
					iHS_resilience	71204669	-2.449172	0
					rsb	70758173	4.9297797	16
					xpehh	70873340	4.8860907	31
		71250000	71500000		iHS_resistance	71015902	-2.575271	0
					iHS_resilience	71204669	-2.449172	0
					rsb	71450224	4.4112068	5
					xpehh	71239666	3.6201579	0
		71500000	71750000		iHS_resistance	71584951	-2.596215	0
					iHS_resilience	71450224	-2.3704	1
					rsb	71450224	4.4112068	3
					xpehh	71328154	3.5225274	0

6.4.5 Gene annotation

In total, 117 genes (69 and 48 in the resilient and resistant lines, respectively) were found in the CNVRs that were unique to the two lines (Table 6.11), therefore these could be responsible for special phenotypes in each line. Besides, the unique CNVRs in resilient and resistant lines were found to overlap with 25 and 24 known parasite related QTL zones, respectively (Table 6.11).

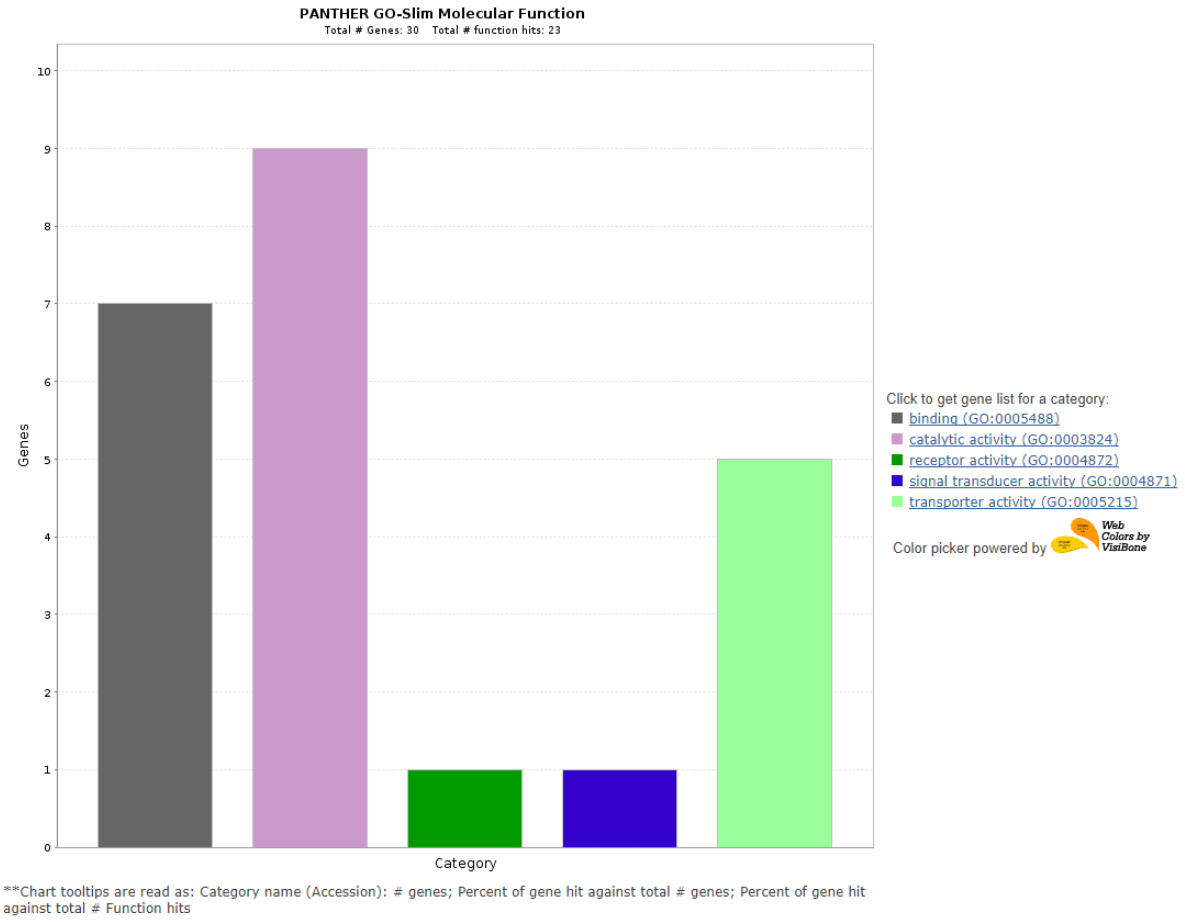


Figure 6.8 Ontology and pathway analysis (molecular function) of genes harbouring the significant SNPs detected by EHH test.

X and Y axis represent the categories and the number of gene hits.

In the SNP study, 13 and 23 genes (Table 6.12) were found in close proximity to the significant (P<0.0001) SNPs detected by EHH test in the resistant and resilient lines, respectively (details in additional file S 6.6) Twenty and thirteen significant SNPs in the resilient and resistant lines were located within 11 and 6 known QTL zones (Table 6.3 and Table 6.4) related to parasite FEC or larva count. Ontology and Pathway analysis showed that these genes are involved in five molecular functions (binding, catalytic activity, receptor activity, signal transducer activity, transporter activity), 11 biological process (biological adhesion, biological regulation, cellular component organization or biogenesis, cellular process, developmental process, immune system process, localization, metabolic process,

multicellular organismal process, reproduction, response to stimulus) and 7 pathways (5-Hydroxytryptamine degredation, cadherin signaling pathway, de novo purine biosynthesis, integrin signalling pathway, PDGF signaling pathway, wnt signaling pathway and p53 pathway) (Figure 6.8, Figure 6.9, Figure 6.10).

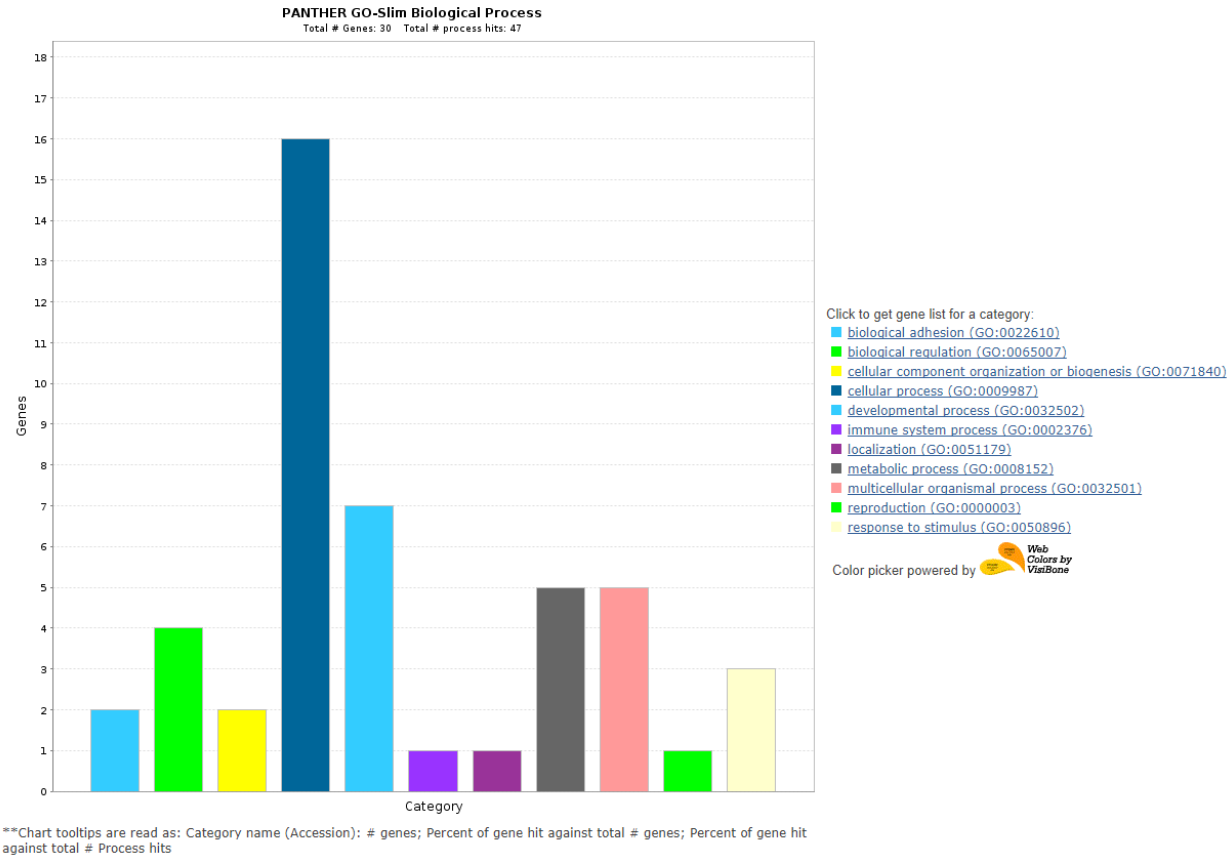


Figure 6.9 Ontology and pathway analysis (biological process) of genes harbouring the significant SNPs detected by EHH test.

X and Y axis represent the categories and the number of gene hits.

Only one gene, PTPRT, was found close to the significant (P<0.0001) SNPs detected by both EHHS tests (Table 6.9). Besides, one and seven genes were found close to the other SNPs detected by XP-EHH and Rsb, respectively (Table 6.6 and Table 6.7). The significant SNPs detected by Rsb and XP-EHH were located within 4 and 2 known parasite related QTL zones, respectively.

Only one SNP (oar3_OAR18_16960784) that was significant only in the Rsb test was located within a unique CNVR found in the resilient line (chr18: 16960478- 16961597). However, there is no gene in this CNVR.

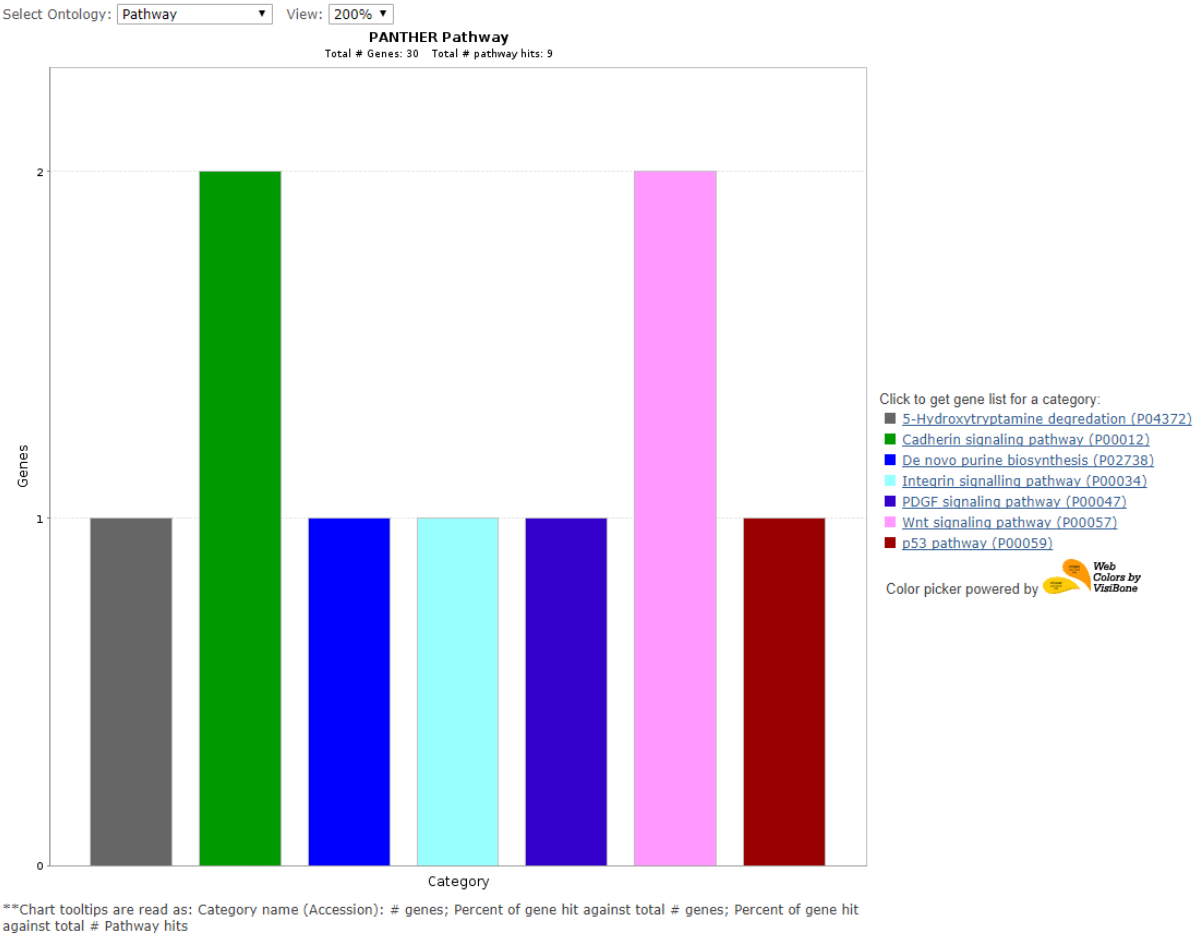


Figure 6.10 Ontology and pathway analysis (pathway) of genes harbouring the significant SNPs detected by EHH test.

X and Y axis represent the categories and the number of gene hits.

Table 6.11 List of genes located within the unique CNVRs, those not common between the two family lines of sheep.

CNVR_ID	Chr	Start	End	line	Gene	QTL	QTL function
CNVR_1	1	7012952	7021637	resilience	ENSOARG00000019204.1		
CNVR_2	1	27035988	27036183	resilience	ENSOARG00000005720.1		
CNVR_5	1	41528974	41530079	resilience			
CNVR_8	1	62312887	62314195	resilience	ENSOARG00000014461.1	QTL:13986	Immunoglobulin A level
CNVR_10	1	86169543	86181785	resilience	ENSOARG00000019297.1		
CNVR_16	1	125889489	125890753	resilience		QTL:13987	Fecal egg count
CNVR_20	1	132887466	132889168	resilience		QTL:12884	Trichostrongylus colubriformis FEC
CNVR_27	1	204625573	204625887	resilience	ENSOARG00000020682.1	QTL:16022	Fecal egg count
CNVR_31_1	1	249738910	249851867	resilience		QTL:12884	Trichostrongylus colubriformis FEC
CNVR_35	1	269500103	269500335	resilience		QTL:12884	Trichostrongylus colubriformis FEC
CNVR_147	2	174777584	174784607	resilience	ENSOARG00000011259.1		
CNVR_149	2	188189636	188195523	resilience		QTL:95615	Immunoglobulin A level
CNVR_152	2	228550985	228553363	resilience	ENSOARG00000020525.1	QTL:19789	Haemonchus contortus FEC
CNVR_207	3	59380272	59399926	resilience	ENSOARG00000020801.1		
CNVR_208	3	59783119	59802100	resilience	ENSOARG00000020835.1		
CNVR_208	3	59783119	59802100	resilience	ENSOARG00000020837.1		
CNVR_210	3	81876613	81879965	resilience			
CNVR_211	3	99670414	99682977	resilience	ENSOARG00000013159.1		
CNVR_212	3	110364598	110364752	resilience	ENSOARG00000016660.1	QTL:12885	Trichostrongylus colubriformis FEC
CNVR_213	3	119351478	119355602	resilience	ENSOARG00000015299.1	QTL:12885	Trichostrongylus colubriformis FEC
CNVR_214	3	132395033	132412540	resilience	ENSOARG00000016342.1		
CNVR_214	3	132395033	132412540	resilience	ENSOARG00000022163.1		
CNVR_214	3	132395033	132412540	resilience	ENSOARG00000023495.1		
CNVR_214	3	132395033	132412540	resilience	ENSOARG00000021214.1		
CNVR_214	3	132395033	132412540	resilience	ENSOARG00000021254.1		
CNVR_214	3	132395033	132412540	resilience	ENSOARG00000016352.1		
CNVR_216	3	138140546	138178721	resilience			
CNVR_217	3	149670559	149673526	resilience	ENSOARG00000020232.1	QTL:12890	Immunoglobulin A level
CNVR_218	3	158425348	158440260	resilience			
CNVR_227	3	204745324	204756705	resilience		QTL:12882	Nematodirus FEC
CNVR_228	3	205615672	205666753	resilience	ENSOARG00000001186.1	QTL:12882	Nematodirus FEC
CNVR_229	3	206712847	206718239	resilience	ENSOARG00000003251.1	QTL:12882	Nematodirus FEC
CNVR_231	3	207312544	207328279	resilience	ENSOARG00000004515.1	QTL:12882	Nematodirus FEC
CNVR_231	3	207312544	207328279	resilience	ENSOARG00000025176.1	QTL:12882	Nematodirus FEC
CNVR_232	3	220421644	220518683	resilience			
CNVR_234	4	49139	74370	resilience			
CNVR_246	4	112222888	112317606	resilience	ENSOARG00000001131.1		
CNVR_246	4	112222888	112317606	resilience	ENSOARG00000001140.1		
CNVR_246	4	112222888	112317606	resilience	ENSOARG00000001183.1		

CNVR_ID	Chr	Start	End	line	Gene	QTL	QTL function
CNVR_246	4	112222888	112317606	resilience	ENSOARG00000001288.1		
CNVR_247	4	112343589	112360805	resilience	ENSOARG00000001140.1		
CNVR_247	4	112343589	112360805	resilience	ENSOARG00000001140.1		
CNVR_247	4	112343589	112360805	resilience	ENSOARG00000001356.1		
CNVR_249	4	112614745	112658570	resilience	ENSOARG00000001537.1		
CNVR_250	4	112742913	112745214	resilience	ENSOARG00000001645.1		
CNVR_254	5	9457037	9458053	resilience	ENSOARG00000005944.1		
CNVR_258	5	20350755	20356584	resilience	ENSOARG00000016187.1		
CNVR_259	5	39181048	39289328	resilience	ENSOARG00000007902.1		
CNVR_259	5	39181048	39289328	resilience	ENSOARG00000007912.1		
CNVR_259	5	39181048	39289328	resilience	ENSOARG00000012049.1		
CNVR_259	5	39181048	39289328	resilience	ENSOARG00000007952.1		
CNVR_259	5	39181048	39289328	resilience	ENSOARG00000012118.1		
CNVR_261	5	46997249	47002753	resilience	ENSOARG00000015873.1		
CNVR_262	5	80768579	80784021	resilience			
CNVR_266	6	2464767	2472560	resilience			
CNVR_267	6	8877236	8877867	resilience	ENSOARG00000000275.1		
CNVR_268	6	10640982	10644670	resilience	ENSOARG00000018069.1		
CNVR_271	6	33543547	33549410	resilience		QTL:16024	Fecal egg count
CNVR_272	6	110465246	110501898	resilience	ENSOARG00000009499.1		
CNVR_273	7	24558	96461	resilience			
CNVR_278	7	24052268	24059426	resilience	ENSOARG00000019796.1	QTL:95614	Fecal egg count
CNVR_278	7	24052268	24059426	resilience	ENSOARG00000013263.1		
CNVR_280	7	32900934	32918695	resilience	ENSOARG00000020143.1		
CNVR_280	7	32900934	32918695	resilience	ENSOARG00000000786.1		
CNVR_283	7	82187154	82193362	resilience	ENSOARG00000000726.1		
CNVR_284	7	85623275	85624732	resilience			
CNVR_289	8	27599054	27607746	resilience	ENSOARG00000010185.1	QTL:12899	Trichostrongylus adult and larva count
CNVR_290	8	42994489	43005714	resilience		QTL:12899	Trichostrongylus adult and larva count
CNVR_292	9	10053530	10055431	resilience		QTL:16026	Fecal egg count
CNVR_293	9	12506894	12653420	resilience		QTL:16026	Fecal egg count
CNVR_296	9	50207112	50209704	resilience	ENSOARG00000026533.1		
CNVR_296	9	50207112	50209704	resilience	ENSOARG00000006401.1		
CNVR_299	9	73131298	73162354	resilience	ENSOARG00000015979.1		
CNVR_38	10	1017882	1044330	resilience			
CNVR_40	10	9642970	9658383	resilience			
CNVR_45	10	50447471	50447906	resilience		QTL:13989	Fecal egg count
CNVR_47	10	59652748	59659908	resilience		QTL:13989	Fecal egg count
CNVR_48	10	70360802	70379733	resilience	ENSOARG00000001021.1	QTL:13989	Fecal egg count
CNVR_49_2	10	70698636	70710040	resilience		QTL:13989	Fecal egg count
CNVR_50	10	71596265	71649636	resilience		QTL:13989	Fecal egg count
CNVR_53_3	10	72067710	72079824	resilience	ENSOARG00000001808.1	QTL:13989	Fecal egg count

CNVR_ID	Chr	Start	End	line	Gene	QTL	QTL function
CNVR_53_4	10	72091260	72113373	resilience	ENSOARG00000001808.1	QTL:13989	Fecal egg count
CNVR_55	10	78781290	78784897	resilience		QTL:13989	Fecal egg count
CNVR_58	11	2819857	2822607	resilience		QTL:12901	Trichostrongylus adult and larva count
CNVR_60	11	5839533	5861956	resilience		QTL:12901	Trichostrongylus adult and larva count
CNVR_64	11	29188972	29194064	resilience		QTL:12901	Trichostrongylus adult and larva count
CNVR_65	11	45401409	45402092	resilience	ENSOARG00000011303.1	QTL:12901	Trichostrongylus adult and larva count
CNVR_69	12	4885725	4910061	resilience	ENSOARG00000008550.1		
CNVR_71	12	35679867	35693556	resilience	ENSOARG00000011451.1		
CNVR_73	13	3870221	3873866	resilience		QTL:95628	Immunoglobulin A level
CNVR_78	13	50676257	50679210	resilience	ENSOARG00000000939.1		
CNVR_79	13	52751652	52754965	resilience			
CNVR_84	14	48888507	48889246	resilience	ENSOARG00000006630.1	QTL:12892	Nematodirus FEC
CNVR_88	14	58781165	58792385	resilience		QTL:12893	Nematodirus FEC
CNVR_95	15	2479390	2501110	resilience			
CNVR_97	15	2994454	3007272	resilience			
CNVR_102	15	67274748	67274984	resilience			
CNVR_106	16	25861474	25861909	resilience			
CNVR_107	16	36140939	36141003	resilience			
CNVR_108	16	47550220	47569525	resilience			
CNVR_111	16	51593643	51594451	resilience			
CNVR_112	16	60474228	60499145	resilience	ENSOARG00000013811.1		
CNVR_115	16	70138435	70139073	resilience			
CNVR_117	17	11079219	11097267	resilience		QTL:16031	Fecal egg count
CNVR_118	17	13782544	13804099	resilience		QTL:16031	Fecal egg count
CNVR_122	17	70144459	70158224	resilience			
CNVR_125	18	16960478	16961597	resilience			
CNVR_126	18	22461162	22486284	resilience		QTL:19806	Worm count
CNVR_127	18	32099419	32100583	resilience	ENSOARG00000001908.1	QTL:19806	Worm count
CNVR_129	18	37236080	37241790	resilience			
CNVR_130	18	45377111	45388471	resilience	ENSOARG00000007986.1		
CNVR_132	18	49442210	49476283	resilience		QTL:12965	Haemonchus contortus FEC
CNVR_133	18	56405850	56406083	resilience	ENSOARG00000013250.1	QTL:12965	Haemonchus contortus FEC
CNVR_134	19	6563666	6563885	resilience			
CNVR_135	19	8026605	8028899	resilience			
CNVR_155	20	1792320	1803350	resilience			
CNVR_157	20	16840728	16843365	resilience		QTL:12896	Immunoglobulin A level
CNVR_161	20	25870527	25876183	resilience		QTL:12896	Immunoglobulin A level
CNVR_162	20	25890124	25910195	resilience			
CNVR_163	20	25985702	25986430	resilience			
CNVR_168	20	28963996	28966049	resilience	ENSOARG00000009041.1		
CNVR_171	20	32005915	32035150	resilience	ENSOARG00000005945.1		
CNVR_171	20	32005915	32035150	resilience	ENSOARG00000005997.1		

CNVR_ID	Chr	Start	End	line	Gene	QTL	QTL function
CNVR_171	20	32005915	32035150	resilience	ENSOARG00000006103.1		
CNVR_174	21	4925532	4941791	resilience			
CNVR_175	21	26199619	26212808	resilience			
CNVR_180	22	15546581	15588904	resilience			
CNVR_183	23	271050	272257	resilience			
CNVR_186	23	11984759	11988113	resilience		QTL:19791	Haemonchus contortus FEC
CNVR_187	23	32665944	32674499	resilience	ENSOARG00000007396.1	QTL:12902	Immunoglobulin G level
CNVR_189	24	36201287	36210968	resilience			
CNVR_191	24	41477084	41485328	resilience	ENSOARG00000024716.1		
CNVR_196	25	21246288	21257915	resilience			
CNVR_200	25	41702268	41711767	resilience	ENSOARG00000001145.1		
CNVR_201	26	19392462	19392969	resilience		QTL:19816	Worm count
CNVR_6	1	81887342	81888299	resistance	ENSOARG00000018785.1		
CNVR_14	1	129311104	129411310	resistance	ENSOARG00000024328.1	QTL:13987	Fecal egg count
CNVR_17	1	132742057	132745876	resistance			
CNVR_19	1	149537506	149543060	resistance			
CNVR_25	1	189420361	189420977	resistance	ENSOARG00000020351.1		
CNVR_29	1	239989502	240013480	resistance		QTL:12884	Trichostrongylus colubriformis FEC
CNVR_30	1	253269290	253269362	resistance		QTL:12884	Trichostrongylus colubriformis FEC
CNVR_32	1	262975317	262976170	resistance		QTL:12884	Trichostrongylus colubriformis FEC
CNVR_134	2	5480897	5492869	resistance			
CNVR_140	2	103366440	103367673	resistance	ENSOARG00000015120.1	QTL:12898	Trichostrongylus adult and larva count
CNVR_148	2	199707868	199749860	resistance			
CNVR_150	2	210669770	210671165	resistance	ENSOARG00000019155.1		
CNVR_151	2	212932871	212940600	resistance			
CNVR_201	3	51174205	51174789	resistance			
CNVR_202	3	53556398	53588042	resistance			
CNVR_203	3	54316198	54328561	resistance			
CNVR_205	3	66694000	66694226	resistance			
CNVR_216	3	204127898	204149926	resistance	ENSOARG00000020983.1	QTL:12882	Nematodirus FEC
CNVR_216	3	204127898	204149926	resistance	ENSOARG00000020985.1	QTL:12882	Nematodirus FEC
CNVR_217	3	204193376	204195677	resistance	ENSOARG00000020995.1	QTL:12882	Nematodirus FEC
CNVR_218	3	205586591	205611089	resistance		QTL:12882	Nematodirus FEC
CNVR_219	3	206982075	206989302	resistance		QTL:12882	Nematodirus FEC
CNVR_222	3	217961445	217972364	resistance			
CNVR_224	4	8760633	8765920	resistance			
CNVR_245	5	49702285	49709967	resistance	ENSOARG00000018844.1		
CNVR_245	5	49702285	49709967	resistance	ENSOARG00000014004.1		
CNVR_249	6	9395636	9396305	resistance			
CNVR_250	6	14385630	14398883	resistance			
CNVR_253	6	35635686	35648306	resistance	ENSOARG00000000411.1		
CNVR_254	6	52726839	52727609	resistance		QTL:16024	Fecal egg count
CNVR_255	6	70027128	70031135	resistance		QTL:13988	Fecal egg count

CNVR_ID	Chr	Start	End	line	Gene	QTL	QTL function
CNVR_262	7	23554581	23558794	resistance	ENSOARG00000012752.1	QTL:95614	Fecal egg count
CNVR_264	7	32433567	32435855	resistance	ENSOARG00000020072.1		
CNVR_266	7	37060296	37071741	resistance	ENSOARG00000020597.1	QTL:95616	Fecal egg count
CNVR_268	7	77542986	77544389	resistance			
CNVR_274	8	73843781	73896231	resistance	ENSOARG00000002907.1	QTL:12899	Trichostrongylus adult and larva count
CNVR_274	8	73843781	73896231	resistance	ENSOARG00000002941.1	QTL:12899	Trichostrongylus adult and larva count
CNVR_275	8	76494970	76507560	resistance	ENSOARG00000003999.1	QTL:12899	Trichostrongylus adult and larva count
CNVR_277	8	82472469	82490641	resistance		QTL:12899	Trichostrongylus adult and larva count
CNVR_281	9	65698425	65709610	resistance		QTL:95623	Immunoglobulin A level
CNVR_283	9	78716461	78733197	resistance	ENSOARG00000003535.1		
CNVR_39	10	7961315	7964322	resistance			
CNVR_42	10	30930742	30931408	resistance		QTL:13989	Fecal egg count
CNVR_45	10	44894527	44895412	resistance		QTL:13989	Fecal egg count
CNVR_47	10	63709386	63711747	resistance		QTL:13989	Fecal egg count
CNVR_49	10	70784492	70787396	resistance		QTL:13989	Fecal egg count
CNVR_50_1	10	70861910	70887739	resistance	ENSOARG00000001163.1	QTL:13989	Fecal egg count
CNVR_50_4	10	71207792	71221008	resistance	ENSOARG00000001232.1	QTL:13989	Fecal egg count
CNVR_50_5	10	71289253	71302255	resistance	ENSOARG00000001282.1	QTL:13989	Fecal egg count
CNVR_56	11	1412875	1419049	resistance		QTL:12901	Trichostrongylus adult and larva count
CNVR_57	11	1545962	1546361	resistance		QTL:12901	Trichostrongylus adult and larva count
CNVR_60	11	5781982	5805535	resistance	ENSOARG00000005660.1	QTL:12901	Trichostrongylus adult and larva count
CNVR_66	11	54281007	54289494	resistance	ENSOARG00000007101.1	QTL:12901	Trichostrongylus adult and larva count
CNVR_69	12	6467201	6479124	resistance			
CNVR_70	12	8418695	8427176	resistance			
CNVR_73	12	42507326	42515366	resistance	ENSOARG00000008942.1		
CNVR_74	12	52585395	52591352	resistance	ENSOARG00000011487.1		
CNVR_75	12	52598575	52613008	resistance	ENSOARG00000011511.1		
CNVR_76	12	78318479	78344760	resistance	ENSOARG00000018514.1		
CNVR_77	13	31703545	31704942	resistance	ENSOARG00000004547.1		
CNVR_84	14	28225845	28235974	resistance		QTL:12893	Nematodirus FEC
CNVR_96	15	23406646	23406918	resistance	ENSOARG00000017719.1	QTL:16029	Fecal egg count
CNVR_97	15	45330938	45347456	resistance	ENSOARG00000017243.1		
CNVR_98	15	45347793	45351621	resistance	ENSOARG00000026824.1		
CNVR_98	15	45347793	45351621	resistance	ENSOARG00000017254.1		
CNVR_102	15	47265541	47270799	resistance	ENSOARG00000019109.1		
CNVR_104	15	51370453	51375462	resistance	ENSOARG00000007994.1	QTL:95630	Immunoglobulin A level
CNVR_105	15	70026469	70029791	resistance			
CNVR_107	15	76859628	76868360	resistance	ENSOARG00000011625.1		
CNVR_110	16	14794674	14795163	resistance	ENSOARG00000025172.1		
CNVR_111	16	20472249	20492290	resistance			
CNVR_112	16	21355830	21377006	resistance			
CNVR_113	16	44703968	44706075	resistance			

CNVR_ID	Chr	Start	End	line	Gene	QTL	QTL function
CNVR_114	16	49624463	49641560	resistance			
CNVR_120	17	26790884	26809383	resistance		QTL:16031	Fecal egg count
CNVR_121	17	35709899	35719269	resistance			
CNVR_125	17	62908484	62913076	resistance	ENSOARG00000014141.1	QTL:95634	Immunoglobulin A level
CNVR_128	18	19088727	19088821	resistance		QTL:19806	Worm count
CNVR_131	18	51981702	51982603	resistance		QTL:12965	Haemonchus contortus FEC
CNVR_132	18	67817610	67852919	resistance	ENSOARG00000007297.1		
CNVR_132	18	67817610	67852919	resistance	ENSOARG00000026475.1		
CNVR_132	18	67817610	67852919	resistance	ENSOARG00000007483.1		
CNVR_132	18	67817610	67852919	resistance	ENSOARG00000007702.1		
CNVR_155	20	11329106	11329475	resistance	ENSOARG00000014556.1		
CNVR_158	20	25685291	25772893	resistance	ENSOARG00000016496.1	QTL:12896	Immunoglobulin A level
CNVR_166	20	50429755	50434588	resistance			
CNVR_172	22	1868181	1893824	resistance		QTL:95609	Immunoglobulin A level
CNVR_173	22	3066890	3074365	resistance		QTL:95609	Immunoglobulin A level
CNVR_178	23	13328865	13329429	resistance		QTL:19791	Haemonchus contortus FEC
CNVR_179	23	17045254	17069800	resistance		QTL:19791	Haemonchus contortus FEC
CNVR_180	23	24931211	24942380	resistance		QTL:95641	Immunoglobulin A level
CNVR_181	23	54207729	54213884	resistance		QTL:95644	Immunoglobulin A level
CNVR_182	24	219096	224492	resistance	ENSOARG00000011293.1		
CNVR_183	24	897190	914967	resistance	ENSOARG00000025873.1		
CNVR_186	24	38259944	38260749	resistance	ENSOARG00000002019.1		
CNVR_186	24	38259944	38260749	resistance	ENSOARG00000002315.1		
CNVR_187	24	40150160	40162060	resistance			
CNVR_192	25	21543868	21552018	resistance			
CNVR_196	26	33079384	33081485	resistance	ENSOARG00000001975.1		

Table 6.12 List of genes located close to the significant (P<0.0001) SNPs detected by EHH test in the gastrointestinal nematode resistant and resilient lines of Romney sheep.

Line	SNP	Chr	Position	Gene
resistance	oar3_OAR1_109497543	1	109497543	ATP1A2
	oar3_OAR1_171568286	1	171568286	HHLA2
	oar3_OAR2_151911086	2	151911086	GPD2
	oar3_OAR2_155966958	2	155966958	FMNL2
	oar3_OAR2_164695219	2	164695219	GTDC1
	oar3_OAR2_165998566	2	165998566	KYNU
	oar3_OAR2_168813588	2	168813588	LRP1B
	oar3_OAR8_82213039	8	82213039	FNDC1
	OAR9_5518009.1	9	5641037	ADGRB3
	oar3_OAR10_79774280	10	79774280	ENSOARG00000005401
	oar3_OAR12_36748585	12	36748585	MROH9
	oar3_OAR17_36538740	17	36538740	FSTL5
	oar3_OAR17_36559268	17	36559268	
	oar3_OAR21_8407961	21	8407961	ME3
resilience	oar3_OAR2_63269803	2	63269803	ALDH1A1
	oar3_OAR2_212143325	2	212143325	ERBB4
	oar3_OAR2_216275404	2	216275404	ATIC
	oar3_OAR3_168208985	3	168208985	ENSOART00000014456
	oar3_OAR4_60004430	4	60004430	ELMO1
	oar3_OAR5_81415565	5	81415565	EDIL3
	oar3_OAR6_16496167	6	16496167	COL25A1
	oar3_OAR10_74818035	10	74818035	STK24
	oar3_OAR11_23841741	11	23841741	ASPA
	s11567.1	14	8824045	CDH13
	oar3_OAR14_24936118	14	24936118	CCDC102A
	oar3_OAR14_42469160	14	42469160	NUDT19
	oar3_OAR14_42723029	14	42723029	RHPN2
	oar3_OAR14_42753314	14	42753314	
	oar3_OAR15_7478160	15	7478160	ARHGAP42
	oar3_OAR15_8612651	15	8612651	CNTN5
	oar3_OAR15_28418100	15	28418100	MPZL3
	oar3_OAR15_28424239	15	28424239	
	oar3_OAR15_28443572	15	28443572	MPZL2
	oar3_OAR15_76299782	15	76299782	ENSOART00000009126
	oar3_OAR17_15472520	17	15472520	INPP4B
	oar3_OAR21_23365859	21	23365859	ENSOART00000008626
	oar3_OAR21_23385917	21	23385917	
	oar3_OAR21_23419044	21	23419044	
	oar3_OAR21_23459252	21	23459252	
	oar3_OAR21_23460758	21	23460758	
	oar3_OAR21_27649204	21	27649204	CCDC15
	oar3_OAR22_41630273	22	41630273	ENSOART00000009611
	oar3_OAR24_24412081	24	24412081	ENSOART00000027850

6.5 Discussion

The two selection lines investigated in the study were selectively bred for 24 years (1985-2009) based on faecal egg count (FEC) BLUP techniques and then randomly bred within each line until the time of sampling in the current study, 2015. This makes these lines ideal to observe the effects of long term artificial selective breeding. The purpose of this study was to compare CNV differences and identify SNPs exhibiting positive selection from the pressure of gastro-intestinal nematodes in two lines of Romney sheep selected for resistance or resilience.

CNV analysis revealed 314 and 294 CNVRs in the resilient and resistant lines respectively. About 48.8 % CNVRs were shared between the lines (Figure 6.5), 182 genes were detected in those regions. There were 114 and 91 CNVRs unique to the resilient and resistant lines containing 69 and 48 genes, respectively, which included olfactory receptor, HLA class II histocompatibility antigen, GTPase IMAP family member, SLA class II histocompatibility antigen, etc. Of these, olfactory receptor coding gene and histocompatibility antigen coding gene as important. In this study, different types of the olfactory receptor gene were found in the resilient and resistant lines. Similar gene has been reported in a former paper (Hou et al. 2012c) where only CNVRs found in parasite-susceptible cattle carried olfactory receptor coding gene. It indicated olfactory receptor coding gene could have an important role in resistance to parasite in ruminants. Besides, histocompatibility antigen coding gene is also important because it has been found to be associated to resistance or resilience of intestinal nematodes in ruminant (Paterson 1998).

On the other hand, several methods based on different theories, such as frequency-based, linkage disequilibrium-based (EHH and EHHS), population differentiation– based (Fst), have been used in different studies to detect selection signatures. However, there is no evidence

show which one is the best (Vitti et al. 2013). In the current study, we employed the EHH and EHHS methods to detect SNP-based selection signatures in the two lines of sheep. Using the EHH test, 62 SNPs showing positive selection signatures were found in the resistant line and 85 in the resilient line. None of those SNP was shared between the two lines, indicating genetic differences between these two lines. Significance of these SNPs could due to other reasons, such as breed, environment, rather than the selection pressure from parasite. In total, 36 genes were found close to these significant SNPs detected in the two lines. Ontology and Pathway analysis showed only one gene, ADGRB3, to be directly related to immune system. This gene encodes a brain-specific angiogenesis inhibitor, a seven-span transmembrane protein, and is thought to be a member of the secretin receptor family and involves p53 pathway which has been reported as an important factor for plasmodium liver-stage infection (Kaushansky et al. 2013). The finding of multi molecular function, biological process and pathways in ontology and pathway analysis indicated the resilience/resistance to nematodes in sheep involved a complicated multi genes or systems interaction which need to be studied further.

Two (FNDC1 and ENSOARG00000005401) out of the 13 genes in the resistant line and 9 (STK24, ASPA, CDH13, CCDC102A, NUDT19, RHPN2, MPZL3, MPZL2, INPP4B) out of the 23 in the resilient line, detected close to the significant SNPs were located in the known parasite-related QTL regions (Additional file: S 6.6). These genes have a variety of functions, ranging from activation of G protein to catalysis the deacetylation of N-acetylaspartic acid. So far, it is hard to give a direct conclusion for the relationship between these genes and resistance\resilience.

Two algorithms, XP-EHH and Rsb were used to detect positive selection signatures between lines. So far, there is no demonstrated advantage of one over the other and the results from

these two algorithms are quite similar (Gautier et al. 2017). In the current study, the results from the two algorithms were mostly similar, there were a few differences with respect to log(*p*) values for the statistics. Figure 6.11 depicts the distribution of log(p) values for Rsb and XP-EHH with respect to markers on chromosome 13, while Figure 6.12 depicts a strong correlation between the log(*p*) values for the two statistics with respect to the markers on the same chromosome. A correlation coefficient of 0.79 was observed between the two statistics, which is close (0.84) to that reported by Gautier et al. (2017). Of the total 39 and 48 SNPs that were detected to exhibit positive selection signatures by the two algorithms respectively in the current study, ten SNPs (Table 6.9) were found to be shared between the two algorithms. Further, the ten SNPs shared by two algorithms were found to be located in regions of known significance (Table 6.10) and hence, can be considered as highly confident SNPs for nematode resistance in sheep. SNP oar3_OAR11_48327544 is located on chromosome 11, within the QTL:12901, known to be associated with nematode larva count (Crawford et al. 2006). The remaining nine SNPs are all located within a small region overlapping the gene Protein Tyrosine Phosphatase, Receptor Type T (PTPRT) on chromosome 13 (Oar_v4.0). Further, these SNPs are harboured within the previously identified QTL:16027, associated with nematode FEC (Silva et al. 2012). *PTPRT* is a protein coding gene and gene ontology annotations show that this gene is related to phosphatase activity and beta-catenin binding, possibly indicating that this gene could have a function in resistance to nematodes. Of the remaining 77 significant SNPs detected by either of the algorithms, 29 SNPs detected by XP-EHH were all were located within the same QTL regions (QTL:16027) (Silva et al. 2012) as above, while 38 SNPs detected by Rsb were located within four previously identified QTLs (QTL:12901 (Crawford et al. 2006), QTL:16027 (Silva et al. 2012), QTL:12965 (Marshall et al. 2009), QTL:12898 (Crawford et al. 2006).

A previous study (McRae et al. 2014) on Romney and Perendale sheep, using F_{ST} and Peddrift to detect differentiation between gastrointestinal nematode resistant and susceptible lines, identified sixteen significant regions, which included candidate genes involved in chitinase activity and the cytokine response. However, none of the SNPs found significant in the current study overlapped with those of McRae et al. (2014). This could be due to two probable reasons. The samples from nematode resistant sheep obtained for the current study came from the same line as that from McRae et al. (2014), but were from a different generation. Besides, the previous study (McRae et al. 2014) used population differentiation–based Fst test, rather than linkage disequilibrium-based EHH and EHHS tests.

Two selective signature regions were found on chromosome13. No gene was found harboured within or near one of the regions, while the other region had 9 genes involved in cell growth, the splicing factor SR family, mitosis, serine/threonine protein kinase, biosynthesis. The result shows that selective signatures tend to gather incertain regions of the chromosome, rather than even distribution, which could because of some kind of unknown mechanism in evolution.

None of the ten SNPs that were significant in both the EHHS tests, Rsb and XP-EHH, are located in the CNVRs identified in the resilient or resistant lines, which indicates that CNV and SNP could represent different aspects of genetics respectively. Only one SNP (oar3_OAR18_16960784) that was significant only in the Rsb test was located within a unique CNVR found in the resilient line (chr18: 16,960,478- 16,961,597). However, there is no gene in this CNVR.

Revisiting the significant SNP and CNV regions subsequent to improvements in the ovine gene map in future might identify a few functional genes located in the regions. Also, cross-species comparison (blasting in Ensembl) of the identified genes could be a good choice for

further study. Besides, only one similar paper pulished in cattle that found the genomic regions associated BC-MFEC (Box-Cox transformed mean FEC) have not been affected by local autozygosity or recent experimental selection.

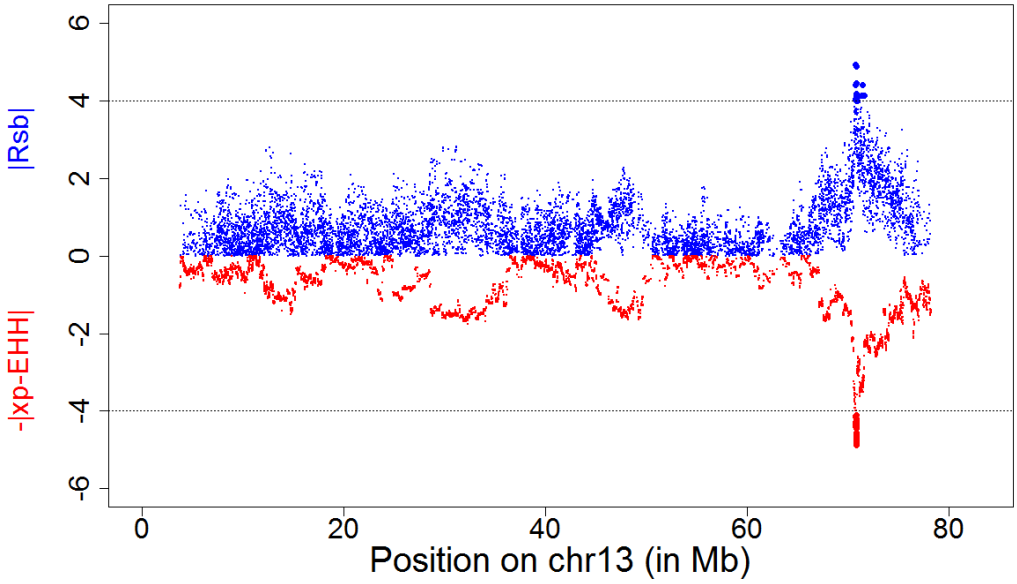


Figure 6.11 Plot showing the distribution of log(*p*) values for Rsb and XP-EHH, between-line site-specific extended haplotype homozygosity (EHHS) test statistics, with regard to single nucleotide polymorphism (SNP) loci located on chromosome 13, in two Romney sheep lines (gastrointestinal nematode resistant and resilient).

The blue and red points represent Rsb and xp-EHH values, respectively. The absolute value on Y axis represents log(p-value). The two grey lines represent log(p-value) thresholds for staistic and the bold points represent the SNPs that passed the threshold (dotted line). X axis represents position on the chromosome

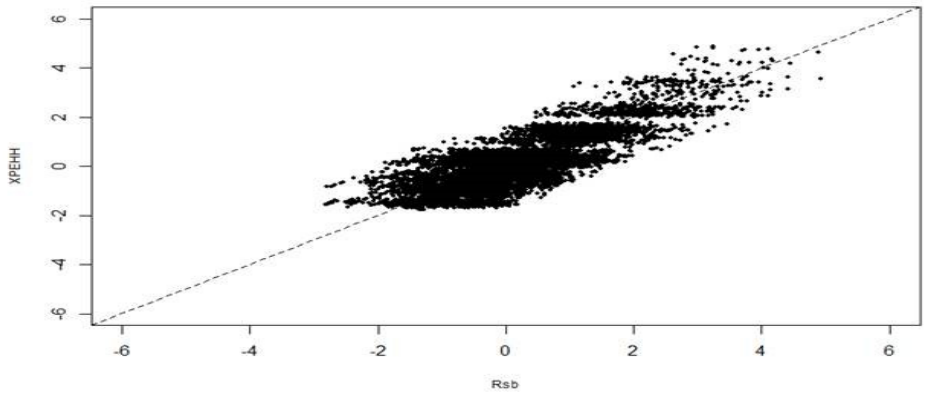


Figure 6.12 Plot showing correlation between XPEHH and Rsb statistics for markers located on chromosome 13
x and y axis represents Rsb and XPEHH. The points represent correlation value of SNPs.

6.6 Conclusion

This study provided a genome-wide map of positive selection signatures in two Romney sheep lines selectively bred for either nematode resistance or resilience, based on FEC. Several significant SNPs were identified in within line EHH and between-line EHHS (XP-EHH and Rsb) tests. Ten SNPs by both XP-EHH and Rsb test were located within two previously detected QTLs associated with gastrointestinal nematodiasis in sheep. By overlapping the areas having significant SNPs, two selection signature regions were detected and one of them contained 9 genes. Besides, only small proportion of the CNVRs detected in the resilient and resistant lines overlapped, indicating huge genetic differences between the lines. Finally, only one SNP detected by Rsb overlapped to a CNVR found in the resilient line, indicating that SNP-based selection signatures and CNV could represent different aspects of sheep immuno-genetics.

6.7 Additional files

Table S6.1 sample information.xlsx

Table S6.2 ihs_resistant_all.xlsx

Table S6.3 ihs_resilient_all.xlsx

Table S6.4 xpehh_all.xlsx

Table S6.5 Rsb_all.xlsx

Table S6.6 Gene annotation.xlsx

Table S6.7 CNV.xlsx

Chapter 7

General discussion

7.1 Thesis objective

The primary objective of this thesis was to explore the utility of genome-wide copy number variation (CNV) as a genetic marker for the analysis of quantitative traits in sheep. To fulfil the objective, five different studies (chapters 2 to 6) were undertaken. The first two studies were aimed at detecting CNV using 50 K SNP BeadChip genotype data (chapter 2) and next generation sequencing (NGS) data (chapter 3). Subsequently, using CNV detected from high density SNP genotype data a genome-wide association study (chapter 4) was undertaken with the aim to identify markers for three phenotypes pertaining to gastrointestinal nematodiasis in sheep. Somatic mosaicism of CNV in adult and foetal tissues was examined inchapter 5. The final study (chapter 6) detected CNV differences and SNP based selection signatures in two Romney lines selected for gastrointestinal nematode resistance or resilience. Results from the five studies have been summarised (7.2) and discussed (7.3).

7.2 Summary of results

Three different algorithms, SVS, PennCNV and cnvPartition were used to detect CNV based on 50K SNP microarray data from 385 sheep belonging to different genetic groups (chapter 2). The results showed large CNV differences among five breeds which suggest that CNVs could be used as genetic marker. However, different CNVs identified by different algorithms, indicating that reliability of CNV calling could be a challenge in the usage of CNVs as genetic markers.

In this study (chapter 3), CNVs were successfully detected in five animals using NGS data. Compared with previous studies in sheep that used either 50 K and/or HD SNP data, or aCGH

showed that NGS provided highest CNVR resolution. The per individual call rate was higher than other platform, so CNVRs were smaller in size. The study showed that a good reference genome and high sequencing depth were essential for efficient CNV detection using NGS. A comparison of pedigree indicated that while most CNVs were inherited in a Mendelian fashion, a few exceptions were seen. This suggested that either spontaneous mutations occurred or the techniques used to detect CNVs were not entirely accurate.

This study (chapter 4) examined association of genome-wide CNV and SNPs with three different phenotypes, live weight, immunity to nematodes and faecal egg count, in Romney sheep. Only three significant CNVRs were associated with live weight and one with FEC. Seven genes were located within the four CNVRs, including an olfactory receptor gene, a gene involved in neuronal differentiation, a putative killer cell immunoglobulin-like receptor and a class II histocompatibility antigen. The olfactory receptor coding gene is interesting because it was previously reported, that only CNVRs found in parasite-susceptible cattle carried olfactory receptor coding gene (Hou et al. 2012c). All four CNVRs overlapped with three previously reported QTL zones, QTL:12891, QTL: 12893, QTL 12896 (Davies et al. 2006). Two CNVRs (QTL:12891 and QTL 12896) were associated with *Nematodirus* spp. FEC and the other (QTL: 12893) with an immunoglobulin A level, which was reported to be associated with resistance/resilience to the nematodirus parasite in sheep (Davies et al. 2006). No SNPs were significant for three kinds traits at the genome- wide scale, probably due to a combination of the very small sample size and their close genetic distance. There was no overlap between CNV- and SNP- based GWAS results; this demonstrated that SNPs and CNVs could be independent of each other.

In this study (chapter 5), somatic mosaicism of CNV explored in adult and foetal tissues using two algorithms, namely cnvPartition and PennCNV. The results showed that significant

CNV mosaicism existed in sheep and that mosaicism was influenced by age (high in foetuses when compared to adults), individuals, the CNV detection algorithm and the type of tissue analysed. Employing a combination of CNV detection algorithms, rather than individual algorithms, will be crucial in order to achieve a sufficiently high accuracy to estimate somatic mosaicism.

The study (chapter 6) provided a genome-wide map of positive selection signatures in two Romney sheep lines selectively bred for either nematode resistance or resilience, based on FEC. Many significant SNPs were identified; ten of them were detected in both EHHS tests (XP-EHH and Rsb), which were located within two previously detected QTLs, namely QTL 12901 and QTL 16027, associated with gastrointestinal nematodiasis in sheep. By overlapping the areas with significant SNPs, two selection signature regions containing nine genes were detected. Only a small proportion of the detected CNVRs overlapped between the resilient and resistant lines, indicating large genetic differences between lines that could possibly contribute to the differential phenotypes in the two lines. Only one significant SNP overlapped to a CNVR in the resilient sheep, indicating SNP selection signatures and CNVs could represent different genetic markers for sheep.

7.3 Discussion of results

7.3.1 Genotyping platform

Two genotyping platforms, SNP microarray and whole genome sequencing, were used to detect CNV in the studies described in this thesis. Their advantages and disadvantages are discussed below.

7.3.1.1 SNP microarray

Overall, the main advantage of the SNP microarray is its low cost and high throughput, with an average price of about $250 per sample which is a good choice for large population

genetics research. In this study, two different densities of SNP microarrays were used, the Illumina OvineSNP50 BeadChips and the Ovine Infinium® HD SNP BeadChip. The main difference between these two microarrays is the number of probes. The former has 54,241 probes (Oddy et al. 2007) while the latter has 606,005 (Kijas et al. 2012). Based on the total genome size for sheep (2,534,344,180 base pairs) (Jiang et al. 2014b), the average gap between successive SNP in the low and high density microarrays are 46,723 and 4,182 bp respectively. This calculation will be biased because the probes are not evenly distributed across the chromosomes. This suggests that on average, the CNVs less than 46,723 bp detected bythe Illumina OvineSNP50 BeadChips or the CNVs less than 4,182 bp detected by the Ovine Infinium® HD SNP BeadChip would not be reliable. It is generally recognised that more CNVs can be detected while more probes are applied. In the current study, the average per-individual CNV detection rate of the OvineSNP50 BeadChips using the PennCNV algorithm was 12.3 (4758 CNVs from 385 samples), which was lower than that of the Ovine Infinium® HD SNP BeadChip (35.6, 3313 CNVs from 93 samples).

There are two inherent deficiencies limiting CNV detection using SNP microarrays, uneven SNP distribution across the chromosomes and gaps between SNPs. Firstly, it is not possible to detect CNVs in regions where no SNPs are present. Secondly, it is hard to find CNVs which are smaller than the size of the gaps between SNPs. Therefore, it is hard to get complete CNV information based on SNP microarrays.

7.3.1.2 Whole genome sequence

Whole genome sequencing (WGS) is a better choice for CNV detection than SNP microarrays. Firstly, WGS enables detection of more CNVs than SNP microarrays. In this study, about 17,000 CNVs were detected before quality control for each sample. However, CNV detection using WGS is highly dependent on the quality of the reference genome. The gaps on the reference genome will result in detection errors, as a consequence the CNVs

overlapped onto these gaps have to be excluded. In the sheep genome version used in these studies (Oar_v3.1), there are 126,619 gaps which will significantly reduce the discovery of CNVs. Thus, the final average CNV per sample after quality control in the current study was 662 CNVs. Even so, the number of CNVs detected by WGS is still much higher than that by SNP microarray. With future improved versions of the ovine reference genome, more CNVs will be retained. In addition, by increasing the depth of sequencing, the number of CNV detected will increase (Chapter 3). The average depth of this study was just 10x which is much less than the current common CNV study (30x). Therefore, it is reasonable to believe that not all CNVs were discovered.

The second advantage of WGS is that it supports better resolution than SNP microarray. The mean and median size of CNVRs detected in this study were 3,835 and 1,999 bp. This overcomes the limitation of CNV detection using SNP microarray (46,723 bp detected in Illumina OvineSNP50 BeadChips or 4,182 bp in the Ovine Infinium® HD SNP BeadChip). WGS also has some challenges. Most important is that read-length is short (150-300 bp), which makes it difficult to map highly repetitive regions to the reference genome. It is also the reason why there are so many gaps in the reference genome. Besides, library preparation for NGS needs an amplification step to get fragments. However, PCR can result in bias for sample composition, leading to potential underrepresentation for low frequency fragments.

7.3.1.3 Inherent platform and sample bias in the current study

This study employed multiple platforms (SNP arrays and NGS) for CNV prediction. Also, two different SNP arrays (OvineSNP50 Ovine Infinium® HD) were used and the SNP genotyping was processed by different companies. Hence, there might be batch effects which could influence the log R ratio intensity, thereby effecting the CNV calling. Besides, DNA was extracted from different tissues and samples and the resultant differences in DNA quality

could also influence CNV results. Further, genetic background information was unavailable for most of the animals geneotyped and hence, pedigree analysis could be undertaken.

7.3.1.4 Conclusion

In summary, both SNP microarray and WGS platforms have advantages and disadvantages. The most important advantage of SNP microarray is the relatively cheaper price, which makes it a good choice for large population studies. However, due to inherent deficiencies of SNP microarrays, the resolution of CNVs detection based on this platform is lower than that based on WGS. On the other hand, WGS provides a much better resolution for CNV detection. However, a higher cost is the main barrier for large population studies, which are very important in animal breeding research. Besides, WGS based CNV detection is highly dependent on the quality of the reference genome. However, the current reference genome, Oar_v3.1, is incomplete with too many gaps which dramatically reduce CNV detection. Besides, some potential biases existed in the current study (batch effect, population difference, platform difference and algorithm difference) and these factors should be considered when planning further research.

7.3.2 CNV detection algorithms

So far, there are more than 12 SNP based algorithms and 30 NGS based algorithms published. Three SNP based algorithms (cnvPartition, PennCNV and Golden Helix's SNP variation suite, SVS) and one NGS based algorithm (CNVnator), were used in this study.

Two of the SNP based algorithms, cnvPartition and PennCNV, work on a similar mechanism, by calculating Log R ratio (LRR) and B allele frequency (BAF) to detect systematic deviation from neighbouring SNPs while Golden Helix's SNP variation suite, uses just LRR. In the current study, only 69 CNVRs (7%) of the total CNVs were shared between all three algorithms, indicating significant differences in detection by these algorithms. This

conclusion also is consistent with other studies in humans. Pinto et al. (2011) compared several algorithms for CNV detection and found that the concordance of CNV calls from the different analytical tools, based the same raw data were less than 50%. Besides, frequency distributions of the sizes of CNVRs showed that SVS tended to find smaller CNVRs while cnvPartition tended to find larger CNVRs, a similar finding also observed in a previous study on CNV detection in cattle using these three algorithms (Xu et al. 2013). qPCR validations showed that cnvPartition had the highest accuracy, perhaps because of the false positive detection rate of large CNVRs being lower.

There are five strategies for CNV detection using NGS data, paired-end mapping, split read, read depth, *de novo* assembly of a genome and a combination of these approaches. However, due to the low depth of NGS data in this study, only CNVnator_v0.3.2 (Abyzov et al. 2011), a read depth based algorithm was able to be used in this study. Results showed that CNVs detected by CNVnator are inherited. Nearly 71% of the total CNVs in the first offspring and 65.9% of the total CNVs in the second offspring could be traced from their parents (chapter 3). This is supported by a study in humans. Legault et al. (2015) compared 4 different sequence-based methods (ERDS, CNVnator, CNVer and Breakdancer) for CNV detection using twins and their parents and found the CNV inheritance rate detected by CNVnator algorithm to be 70%, which is very close to that seen in this study. The accuracy of CNV detection using CNVnator was hard to validate in this study, because DNA from only one tissue sample could be extracted due to its long-term storage. However, Legault et al. (2015) used monozygotic twins sharing the same genome as a comparison to estimate the error rate of different algorithms and found the accuracy of CNVnator to be good.

In summary, large differences exist in CNVs detected by different algorithms. Therefore, evaluation of the accuracy of each algorithm is crucial for CNV data to be used in further

analyses such as phenotype association studies. However, in the case of sheep as well as other domestic animals, there is no gold standard of CNV calls to compare data against. Hence, it would be meaningful to employ multiple algorithms, rather than just one for CNV detection and to use the pooled data for further applications.

7.3.3 CNV validation by qPCR

The qPCR validation of CNVs has its associated challenges. There are two parts discussed below.

7.3.3.1 The purpose of validation

A few CNV studies (Liu et al. 2013; Ma et al. 2015a) validated CNVR instead of CNV. This is not an accurate validation since CNV is a kind of genetic marker like SNP, but CNVR is an integrated value by merging CNVs of the whole population. Therefore, in this study, CNVs, rather than CNVRs were validated by qPCR, as done in a CNV study in humans (Redon et al. 2006).

7.3.3.2 Method of validation

There are different ways to validate CNVs, qPCR being the most popular one. Due to limited financial resources, the relative qPCR (SYBR Green) method was chosen for CNV validations in the current study, also employed in several other studies (Liu et al. 2013; Ma et al. 2015a; Wang et al. 2013a). However, there were some issues regarding this approach. Firstly, CNV normally happens randomly which means it is hard to know whether a particular gene has CNV in the whole population. In this study, the *DGAT1* gene was employed as a control gene because it is considered not to contain any CNV (Fontanesi et al. 2011). It is very difficult to make this kind of conclusion because it needs large-population screening to be sure.

Secondly, the choice of a control sample also is a challenge. In a diploid organism, an ideal control sample should have no CNV (2 copies). However, in order to know which sample is ideal another control sample is needed; therefore, this becomes a logical paradox. In order to overcome this problem, the copy number of the control sample must be assumed. In this study, the control was assumed to have either one, two or three copies. Four copies are unusual and no copies means that there is no gene in the sample being tested therefore there should be no qPCR result at all. Based on the assumed copy number, three kinds of theoretical thresholds were obtained. By comparing qPCR results to these theoretical thresholds, the assumed copy number with the highest accuracy was considered as the correct copy number of the control sample. Of course, this approach may have a bias because the assumed copy number with highest accuracy might not mean that it is true.

Thirdly, the boundary of the threshold was not clear. For example, suppose the copy number at a given locus of the control sample to be 2, therefore the theoretical threshold of copy number variation is 0 for 0 copy, 0.5 for 1 copy, 1 for 2 copies, 1.5 for 3 copies, 2 for 4 copies. However, if the observed value is 0.75, it is hard to know whether the copy number is 1 or 2.

Fourthly, there were systemic errors made because of the SYBR Green technique. All qPCRs using SYBR Green can only test one target gene in one tube at one time. This means the target gene and reference gene of the same sample have to be run in two separate wells or tubes. It is hard to make the template for two tubes absolutely the same which will influence the final result of the CNV validation. Even a slightly higher quantity template in a well will result in false positive gain and slightly less quantity will lead to false positive loss.

7.3.3.3 Conclusion

In summary, so far almost all CNV studies used relative qPCR for CNV validations, since this method is relatively cheaper. However, this kind of validation has some inherent defects because it is hard to select a suitable control gene and control sample as reference. Besides, the boundaries of threshold for different CNV values are not clear and systemic errors of the SYBR Green technique could lead to inaccurate results. Therefore, in order to obtain a reliable validation result, absolute qPCR might be helpful.

7.3.4 CNV as a genetic marker

7.3.4.1 Accuracy of CNV detection

In order to use CNVs as a genetic marker, the first issue to solve is the need to increase the accuracy of CNV detection. So far, the accuracy of CNV detection is not very high. In this study, there were large differences in CNV detection between the chosen platforms and algorithms. Without reliable results, any analysis based on those cannot be trusted. However, due to the high cost of qPCR based validation, it is not possible to test all CNVs to calculate the false positive rate of each algorithm. Therefore, it is necessary to find an alternative method for comparing algorithms. A recent study by Legault et al, that used WGS data from monozygotic twins to estimate the false positive error as a method to compare CNV detection algorithms, is a step in this direction (Legault et al. 2015). A similar approach could be used to select algorithms for SNP-based CNV detection.

7.3.4.2 CNV characteristics

So far, the genetic characteristics of CNV are unclear. In this study, CNVs showed major differences between breeds which indicates CNV could be a useful genetic marker. The pedigree comparison in chapter 3 demonstrated about nearly 70% of CNVs in two half-sibs were inherited from the parents, meaning CNV could be highly inherited. The remainder of the CNVs in the two progenies could be because of random mutation. There might be two

main reasons responsible for this issue. Firstly, if the CNV length is too long, it could be broken due to synapsis. Secondly, the CNV could be a kind of somatic mosaicism so that it will not be inherited by the next generation. Results in Chapter 5 revealed CNV somatic mosaicism to be a common phenomenon, with an overall mosaicism of 58.5% and 42.6% in foetuses and adults, respectively. Also, out of the CNVs inherited by the two half-sibs (106 and 133 CNVs, respectively, by 828-05-01 and 828-05-03), exclusively from the sire, 26 CNVs overlapped (Figure 3.10), further indicating the Mendelian pattern of inheritance of CNV.

7.3.4.3 Conclusion

In summary, CNV could potentially be a useful genetic marker. However, the accuracy of detection needs to be improved, while further work needs to be undertaken to fully understand the genetic characteristics, such as inheritance and mosaicism, of CNV.

7.3.5 GWAS and selection signatures

A GWAS study (chapter 4) was undertaken using both CNV and SNP. The results of CNV and SNP associations did not overlap, indicating that CNV and SNP represent different kinds of genetic markers. This finding is consistent with the observation in a human study that SNPs and CNVs captured 83.6% and 17.7% of the total detected genetic variation in gene expression, respectively, with little overlap (Stranger et al. 2007). However, the GWAS results of the current study could be biased as there were some limitations in the study. Firstly, the small number of samples limited the ability to find significant markers. Besides, all the samples were collected from two sub groups of the same population, with the IBD between them being higher than 0.25. This meant the animals were closely related and this potentially decreased the opportunity to identify significant CNVs or SNPs by GWAS.

In the selection signature study (chapter 6), several line-specific SNP signatures as well unique CNVRs in the gastrointestinal resilient and resistant lines were found. None of the significant SNPs (except one) were within the detected CNVRs, further substantiating the hypothesis that the contributions of SNPs and CNVs to phenotype could be non-overlapping.

7.4 Suggestions for further research

In summary, due to the higher resolution of CNV detection, it can be proposed that NGS will become more popular in CNV-based research than SNP microarray. In order to overcome the drawbacks (such as read length) of NGS, long range sequencing (such as via PacBio or Oxford Nanopore) might be necessary, because its long read (average > 10,000 bp, some reads > 60,000 bp) will dramatically reduce the gaps on reference genomes and increase the mapping rate. In addition, long-range sequencing does not need PCR so that it will not under-represent low proportion sequence. Finally, long range sequencing (third generation sequencing) has no mapping bias unlike the second generation sequencing and the random mismatch can be corrected by increased depth. So far, at least one paper has been published based on the PacBio technique, in cattle with a third generation sequencing based algorithm called Sniffles (Couldrey et al. 2017).

Relative qPCR based on SYBR Green is not the best choice for CNV validation. A few modifications could be applied to improve the quality of validation. Firstly, it is a good idea to use absolute qPCR, instead of relative qPCR. By using absolute qPCR, it is clear to know the absolute copy number of target gene and control gene of a sample. In diploids, if the control gene is reliably known to have 2 copies in species, the quotient of copy number of target gene by control gene will indicate the real copy number of the target sample. Also, in order to eliminate systemic error due to the differences in template, the best way is to run the qPCR of the target gene and the control gene in the same well or tube. Taqman™ by Applied Biosystems (Thermo Fisher Scientific corporation) is a good choice, because the fluorescent

probes of Taqman can carry different colours so that the fluorescent intensity signals of target gene and control gene can be recorded at the same time. A recent study employed this technique for CNV validation in humans (Wang et al. 2016).

Also, a strategy to improve the accuracy of CNV detection is to reduce the false positive results by overlapping the results obtained from different algorithms as suggested by Winchester et al (2009). However, this might also filter out some correct results as a side effect. Therefore, it is also a compromised choice, but not perfect. Finally, machine learning has become popular in various scientific and industrial areas in recent years and will also play an important role in facilitating and speeding up CNV based research. Ding et al. (2014) tried to predict CNV based cancer risk by machine learning.

Finally, a golden CNV method could be deleveloped if many aspects mentioned above improved, such as new platform, new algorithm, new PCR, with a large number experiments.

7.5 Overall conclusion

This thesis explored CNV detection methods, CNV polymorphism as well as its utility as a genetic marker for quantitative traits in sheep. Two kinds of genotyping platforms (SNP microarray and NGS) and four kinds of CNV detection algorithms (SVS, Penncnv, cnvPartition, CNVnator) were tested. Large differences in CNV were evident between genotyping platforms, detection algorithms, breeds and somatic tissues, within individuals.

A CNV-and SNP-based GWAS in sheep selectively bred for resistance or resilience to gastrointestinal nematodes, identified three and one CNVRs to be significantly associated with the phenotypes live weight and FEC, respectively. No SNPs were found to be significant at genome-wide scale. Similarly, probing for selection signatures in those sheep revealed two significant regions (involving 10 SNPs). Several genetic line-specific CNVRs were detected, but only one significant SNP detected in selection signatures overlapped to a CNVR in the

resilient sheep, thus indicating CNVs and SNPs might represent different aspects of immuno-genetics.

Overall, CNV could be a potential genetic marker because it represents completely different genetic information, which will expand understanding of current genetics and support new ideas for genetic breeding. CNV found to be associated with different quantitative triats in livestock can be employed by breeders as reference genetic markers for selective breeding. Besides, they could also provide a clue to reveal potential mechanisms of biochemical and immunological pathways underlying the triats, in further studies of molecular and cell biology. However, the accuracy of CNV detection is still not high enough (95% is the threshold for SNP detection) which increases the uncertainty of CNV-based studies. Besides, genetic characteristics of CNV are not fully known, thus making it hard to develop genetics and statistical models for further analysis. Therefore, future research on CNV should focus on improvement of detection accuracy using new genotyping platforms (such as long-range sequencing) and new machine learning algorithms so as to obtain meaningful results.

References

Abel HJ, Duncavage EJ, Becker N, Armstrong JR, Magrini VJ, Pfeifer JD: SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. Bioinformatics 2010, 26(21):2684-2688.

Abernathy J, Li X, Jia X, Chou W, Lamont SJ, Crooijmans R, Zhou H: Copy number variation in Fayoumi and Leghorn chickens analyzed using array comparative genomic hybridization. Animal Genetics 2014, 45(3):400-411.

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

Abyzov A, Urban AE, Snyder M, Gerstein M: CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 2011, 21.

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E: Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 2006, 439(7078):851-855.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. Genome research 2002, 12(12):1805-1814.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O et al: Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics 2009, 41.

Alonso A, Julià A, Tortosa R, Canaleta C, Cañete JD, Ballina J, Balsa A, Tornero J, Marsal S: CNstream: A method for the identification and genotyping of copy number polymorphisms using Illumina microarrays. BMC bioinformatics 2010, 11:264-264.

Anderson R, Auvray B, Pickering NK, Dodds KG, Bixley MJ, Hyndman D, McEwan JC: Development and characterisation of a low density (5K)

Anderson R: Development of a high density (600K) Illumina Ovine SNP chip and its use to fine map the yellow fat locus. In: Plant and Animal Genome XXII Conference: 2014: Plant and Animal Genome; 2014.

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

Bahbahani H, Salim B, Almathen F, Al Enezi F, Mwacharo JM, Hanotte O: Signatures of positive selection in African Butana and Kenana dairy zebu cattle. PLoS ONE 2018, 13(1):e0190446.

Baker R, Watson T, Bisset S, Vlassoff A: Breeding Romney sheep which are resistant to gastro-intestinal parasites. In: Proc Aust Assoc Anim Breed Genet: 1990; 1990: 173-178.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: Natural selection has driven population differentiation in modern humans. Nature genetics 2008, 40(3):340.

Beckmann J, Soller M: Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. Theoretical and Applied Genetics 1983, 67(1):35-43.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al: Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008, 456(7218):53-59.

Berglund J, Nevalainen EM, Molin AM, Perloski M, Andre C, Zody MC, Sharpe T, Hitte C, Lindblad-Toh K, Lohi H et al: Novel origins of copy number variation in the dog genome. Genome Biology 2012, 13(8).

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF: Copy number variation of individual cattle genomes using next-generation sequencing. Genome research 2012, 22(4):778-790.
Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, Song J, Garcia JF, Sonstegard TS, Van Tassell CP: Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Research 2016, 23(3):253-262.
Biesecker LG, Spinner NB: A genomic view of mosaicism and human disease. Nature Reviews Genetics 2013, 14(5):307-320.
Biosystems A: Guide to performing relative quantitation of gene expression using real-time quantitative PCR. Applied Biosystems, Foster City 2004:28-30.
Biswas S, Akey JM: Genomic insights into positive selection. TRENDS in Genetics 2006, 22(8):437-446.
Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M: Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics 2010, 186(1):241-262.
Botstein D, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics 1980, 32(3):314.
Boussaha M, Esquerré D, Barbieri J, Djari A, Pinton A, Letaief R, Salin G, Escudié F, Roulet A, Fritz S: Genome-wide study of structural variants in bovine Holstein, Montbéliarde and Normande dairy breeds. PloS one 2015, 10(8):e0135931.
Boveri T: The origin of malignant tumors: Williams & Wilkins; 1929.
Brack C, Hirama M, Lenhard-Schuller R, Tonegawa S: A complete immunoglobulin gene is created by somatic recombination. Cell 1978, 15(1):1-14.
Bras J, Guerreiro R, Hardy J: Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease. Nature Reviews Neuroscience 2012, 13(7):453-464.
Bridges CB: The bar" gene" a duplication. Science 1936, 83(2148):210-211.
Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Stahl T, Menzel U, Sandgren J, von Tell D, Poplawski A et al: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 2008, 82(3):763-771.
Campbell IM, Shaw CA, Stankiewicz P, Lupski JR: Somatic mosaicism: implications for disease and transmission genetics. Trends in Genetics 2015, 31(7):382-392.
Carter NP: Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics 2007, 39(7 Suppl):S16-21.
Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015, 4(1):7.
Chen CY, Qiao RM, Wei RX, Guo YM, Ai HS, Ma JW, Ren J, Huang LS: A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. BMC genomics 2012, 13.
Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 2009, 6(9):677.
Chen WK, Swartz JD, Rush LJ, Alvarez CE: Mapping DNA structural variation in dogs. Genome Research 2009, 19(3):500-509.
Chen Y, Liu YJ, Pei YF, Yang TL, Deng FY, Liu XG, Li DY, Deng HW: Copy Number Variations at the Prader-Willi Syndrome Region on Chromosome 15 and associations with Obesity in Whites. Obesity 2011, 19(6):1229-1234.

Chiang DY, Getz G, Jaffe DB, O'kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES: High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature methods 2009, 6(1):99.
Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N: Genome-wide mapping with biallelic markers in Arabidopsis thaliana. Nature genetics 1999, 23(2):203-207.
Christensen B: CNV Analysis Tips for Illumina Data. 2010.
Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P, Nardone A: Massive screening of copy number population-scale variation in Bos taurus genome. BMC genomics 2013, 14.
Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: Basic statistical analysis in genetic case-control studies. Nature protocols 2011, 6(2):121-133.
Clayton D: snpStats: SnpMatrix and XSnpMatrix classes and methods. R package 2012.
Clevert D-A, Mitterecker A, Mayr A, Klambauer G, Tuefferd M, Bondt AD, Talloen W, Göhlmann H, Hochreiter S: cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate. Nucleic acids research 2011, 39(12):e79-e79.
Clop A, Vidal O, Amills M: Copy number variation in the genomes of domestic animals. Anim Genet 2012, 43(5):503-517.
Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J: QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic acids research 2007, 35(6):2013-2025.
Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, Deardorff MA, Krantz ID, Hakonarson H, Spinner NB: Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Human molecular genetics 2010, 19(7):1263-1275.
Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK: A high-resolution survey of deletion polymorphism in the human genome. Nature genetics 2005, 38(1):75-81.
Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J: An estimate of unique DNA sequence heterozygosity in the human genome. Human genetics 1985, 69(3):201-205.
Couldrey C, Keehan M, Johnson T, Tiplady K, Winkelman A, Littlejohn M, Scott A, Kemper K, Hayes B, Davis S: Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. Journal of dairy science 2017, 100(7):5472-5478.
Crawford AM, Paterson KA, Dodds KG, Tascon CD, Williamson PA, Thomson MR, Bisset SA, Beattie AE, Greer GJ, Green RS: Discovery of quantitative trait loci for resistance to parasitic nematode infection in sheep: I. Analysis of outcross pedigrees. BMC genomics 2006, 7(1):178.
Crooijmans RP, Fife MS, Fitzgerald TW, Strickland S, Cheng HH, Kaiser P, Redon R, Groenen MAM: Large scale variation in DNA copy number in chicken breeds. BMC genomics 2013, 14(398).
da Silva JM, Giachetto PF, da Silva LO, Cintra LC, Paiva SR, Yamagishi MEB, Caetano AR: Genome-wide copy number variation (CNV) detection in Nelore cattle reveals highly frequent variants in genome regions harboring QTLs affecting production traits. BMC genomics 2016, 17(1):454.
Darvishi K: Application of Nexus copy number software for CNV detection and analysis. Current protocols in human genetics 2010:4.14. 11-14.14. 28.
Dathe K, Kjaer KW, Brehm A, Meinecke P, Nurnberg P, Neto JC, Brunoni D, Tommerup N, Ott CE, Klopocki E et al: Duplications involving a conserved regulatory element

downstream of BMP2 are associated with brachydactyly type A2. Am J Hum Genet 2009, 84(4):483-492.
Davies G, Stear M, Benothman M, Abuagob O, Kerr A, Mitchell S, Bishop S: Quantitative trait loci associated with parasitic infection in Scottish blackface sheep. Heredity 2006, 96(3):252-258.
de Simoni Gouveia JJ, da Silva MVGB, Paiva SR, de Oliveira SMP: Identification of selection signatures in livestock species. Genetics and Molecular Biology 2014, 37(2):330-342.
de Simoni Gouveia JJ, Paiva SR, McManus CM, Caetano AR, Kijas JW, Facó O, Azevedo HC, de Araujo AM, de Souza CJH, Yamagishi MEB: Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds. Livestock Science 2017, 197:36-45.
Dermitzakis ET, Stranger BE: Genetic variation in human gene expression. Mammalian genome 2006, 17(6):503-508.
Ding X, Tsang S-Y, Ng S-K, Xue H: Application of Machine Learning to Development of Copy Number Variation-based Prediction of Cancer Risk. Genomics Insights 2014, 7:1-11.
Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K: Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic acids research 2008, 36(19):e126.
Doan R, Cohen N, Harrington J, Veazy K, Juras R, Cothran G, McCue ME, Skow L, Dindot SV: Identification of copy number variants in horses. Genome Research 2012, 22(5):899-907.
Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV: Whole-Genome Sequencing and Genetic Variant Analysis of a Quarter Horse Mare. BMC genomics 2012, 13.
Dolezal MA, Bagnato A, Schiavini F, Santus E, Holm L-E, Bendixen C, Panitz F: Copy Number Variation in Brown Swiss Dairy Cattle. In: 10th World Congress on Genetics Applied to Livestock Production (WCGALP): 2014; 2014.
Dong K, Pu Y, Yao N, Shu G, Liu X, He X, Zhao Q, Guan W, Ma Y: Copy number variation detection using SNP genotyping arrays in three Chinese pig breeds. Anim Genet 2015.
Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES: A genetic linkage map of the human genome. Cell 1987, 51(2):319-337.
Dupuis MC, Zhang Z, Durkin K, Charlier C, Lekeux P, Georges M: Detection of copy number variants in the horse genome and examination of their association with recurrent laryngeal neuropathy. Animal Genetics 2013, 44(2):206-208.
Durán Aguilar M, Román Ponce S, Ruiz López F, González Padilla E, Vásquez Peláez C, Bagnato A, Strillacci M: Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. Journal of Animal Breeding and Genetics 2017, 134(1):49-59.
Edwards JH: Familiarity, recessivity and germline mosaicism. Annals of Human Genetics 1989, 53(1):33-47.
Elferink MG, Vallée AA, Jungerius AP, Crooijmans RP, Groenen MA: Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. BMC genomics 2008, 9(1):391.
Fadista J, Nygaard M, Holm L-E, Thomsen B, Bendixen C: A snapshot of CNVs in the pig genome. PLoS One 2008, 3(12):e3916.
Fadista J, Thomsen B, Holm L-E, Bendixen C: Copy number variation in the bovine genome. BMC genomics 2010, 11(1):284.

Fadista J, Manning AK, Florez JC, Groop L: The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics 2016, 24(8):1202.
Familton A, McAnulty R: Life cycles and development of nematode parasites of ruminants. Sustainable control of internal parasites in ruminants 1997:67-80.
Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI: FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nature genetics 2007, 39(6):721-723.
Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B: Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 2013, 193(3):929-941.
Fariello M-I, Servin B, Tosser-Klopp G, Rupp R, Moreno C, San Cristobal M, Boitard S, Consortium ISG: Selection signatures in worldwide sheep populations. PLoS One 2014, 9(8):e103813.
Fernandez AI, Barragan C, Fernandez A, Rodriguez MC, Villanueva B: Copy number variants in a highly inbred Iberian porcine strain. Animal Genetics 2014, 45(3):357-366.
Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. Nature Reviews Genetics 2006, 7(2):85-97.
Fisher RA: Statistical methods for research workers: Genesis Publishing Pvt Ltd; 1925.
Fisher RA: The design of experiments: Oliver And Boyd; Edinburgh; London; 1937.
Fleischmann RD, Adams MD, White O, Clayton RA: Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995, 269(5223):496.
Fontanesi L, Beretti F, Martelli P, Colombo M, Dall'Olio S, Occidente M, Portolano B, Casadio R, Matassino D, Russo V: A first comparative map of copy number variations in the sheep genome. Genomics 2011, 97(3):158-165.
Fontanesi L, Beretti F, Riggio V, Gómez González E, Dall'Olio S, Davoli R, Russo V, Portolano B: Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. Cytogenetic and genome research 2009, 126(4):333-347.
Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, Casadio R, Russo V, Portolano B: An initial comparative map of copy number variations in the goat (Capra hircus) genome. BMC genomics 2010, 11(1):639.
Forsberg LA, Rasi C, Razzaghian HR, Pakalapati G, Waite L, Thilbeault KS, Ronowicz A, Wineinger NE, Tiwari HK, Boomsma D: Age-related somatic structural changes in the nuclear genome of human blood cells. The American Journal of Human Genetics 2012, 90(2):217-228.
Fowler KE, Pong-Wong R, Bauer J, Clemente EJ, Reitter CP, Affara NA, Waite S, Walling GA, Griffin DK: Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. BMC genomics 2013, 14.
Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM: A second generation human haplotype map of over 3.1 million SNPs. Nature 2007, 449(7164):851-861.
Freed D, Stevens EL, Pevsner J: Somatic mosaicism in the human genome. Genes 2014, 5(4):1064-1094.
Gai X, Perin JC, Murphy K, O'Hara R, D'Arcy M, Wenocur A, Xie HM, Rappaport EF, Shaikh TH, White PS: CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. BMC bioinformatics 2010, 11:74-74.

Gautier M, Klassmann A, Vitalis R: rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Molecular Ecology Resources 2017, 17(1):78-90.
Gautier M, Naves M: Footprints of selection in the ancestral admixture of a New World Creole cattle breed. Molecular Ecology 2011, 20(15):3128-3143.
Gill DE, Chao L, Perkins SL, Wolf JB: Genetic mosaicism in plants and clonal animals. Annual Review of Ecology and Systematics 1995:423-444.
Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol L, Schinzel A, Estivill X: Genomic inversions of human chromosome 15q11–q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Human molecular genetics 2003, 12(8):849-858.
Giuffra E, Evans G, Törnsten A, Wales R, Day A, Looft H, Plastow G, Andersson L: The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. Mammalian Genome 1999, 10(12):1132-1136.
Giuffra E, Törnsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JM, Anderson SI, Archibald AL, Andersson L: A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. Mammalian Genome 2002, 13(10):569-577.
Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ: The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005, 307(5714):1434-1440.
Gorla E, Cozzi MC, Román-Ponce SI, Ruiz López FJ, Vega-Murillo VE, Cerolini S, Bagnato A, Strillacci MG: Genomic variability in Mexican chicken population using copy number variants. BMC Genetics 2017, 18(1):61.
Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, Crooijmans RP, Groenen MA, Deryusheva S, Gaginskaya E, Carré W: Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. BMC genomics 2008, 9(1):168.
Grodzicker T, Williams J, Sharp P, Sambrook J: Physical mapping of temperature-sensitive mutations of adenoviruses. In: Cold Spring Harbor symposia on quantitative biology: 1974: Cold Spring Harbor Laboratory Press; 1974: 439-446.
Gurgul A, Jasielczuk I, Szmatoła T, Pawlina K, Ząbek T, Żukowski K, Bugno-Poniewierska M: Genome-wide characteristics of copy number variation in Polish Holstein and Polish Red cattle using SNP genotyping assay. Genetica 2015, 143(2):145-155.
Guryev V, Saar K, Adamovic T, Verheul M, Van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD: Distribution and functional impact of DNA copy number variation in the rat. Nature genetics 2008, 40(5):538-545.
Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S: Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics 2011, 28(1):40-47.
Guðmundsdóttir ÓÓ: Genome-wide association study of muscle traits in Icelandic sheep. 2015.
Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC: Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 2010, 26(10):1277-1283.
Hamada H, Petrino MG, Kakunaga T: A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proceedings of the National Academy of Sciences 1982, 79(21):6465-6469.

Han RL, Yang PK, Tian YD, Wang DD, Zhang ZX, Wang LL, Li ZJ, Jiang RR, Kang XT: Identification and functional characterization of copy number variations in diverse chicken breeds. BMC genomics 2014, 15.
Handsaker RE, Korn JM, Nemesh J, McCarroll SA: Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature genetics 2011, 43(3):269.
Hastings PJ, Lupski JR, Rosenberg SM, Ira G: Mechanisms of change in gene copy number. Nature reviews Genetics 2009, 10(8):551-564.
Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R: InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC bioinformatics 2015, 16(1):169.
Helix G: CNV Univariate Analysis Tutorial. SNP & Variation Suite Manual v8.6. In.; 2017.
Henrichsen CN, Chaignat E, Reymond A: Copy number variants, diseases and gene expression. Human molecular genetics 2009, 18(R1):R1-R8.
Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, Ruedi M, Kaessmann H, Reymond A: Segmental copy number variation shapes tissue transcriptomes. Nature genetics 2009, 41(4):424-429.
Hillel J, Schaap T, Haberfeld A, Jeffreys A, Plotzky Y, Cahaner A, Lavi U: DNA fingerprints applied to gene introgression in breeding programs. Genetics 1990, 124(3):783-789.
Hoeijmakers JH: DNA damage, aging, and cancer. New England Journal of Medicine 2009, 361(15):1475-1485.
Hofker M, Skraastad M, Bergen A, Wapenaar M, Bakker E, Millington-Ward A, van Ommen G, Pearson P: The X chromosome shows less genetic variation at restriction sites than the autosomes. American journal of human genetics 1986, 39(4):438.
Hollox EJ, Hoh B-P: Human gene copy number variation and infectious disease. Human genetics 2014, 133(10):1217-1233.
Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC: Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 2010, 26(12):i350-i357.
Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, Sahinalp SC: Simultaneous structural variation discovery among multiple paired-end sequenced genomes. Genome research 2011, 21(12):2203-2212.
Hou C-L, Meng F-H, Wang W, Wang S-Y, Xing Y-P, Cao J-W, Wu K-F, Liu C-X, Zhang D, Zhang Y-R et al: Genome-wide analysis of copy number variations in Chinese sheep using array comparative genomic hybridization. Small Ruminant Research 2015, 128:19-26.
Hou Y, Bickhart DM, Chung H, Hutchison JL, Norman HD, Connor EE, Liu GE: Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. Functional & integrative genomics 2012, 12(4):717-723.
Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim E-s, Matukumalli LK, Ventura M, Song J, VanRaden PM: Genomic characteristics of cattle copy number variations. BMC genomics 2011, 12(1):127.
Hou Y, Liu GE, Bickhart DM, Matukumalli LK, Li C, Song J, Gasbarre LC, Van Tassell CP, Sonstegard TS: Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. Functional & integrative genomics 2012, 12(1):81-92.
Hou YL, Bickhart DM, Hvinden ML, Li CJ, Song JZ, Boichard DA, Fritz S, Eggen A, DeNise S, Wiggans GR et al: Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. BMC genomics 2012, 13.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nature genetics 2004, 36(9):949-951.
Illumina, Illumina ovine SNP chip. In: Plant & Animal Genomes XX. San Diego, CA, United States of America.; 2012.
illumina: DNA Copy Number and Loss of Heterozygosity Analysis Algorithms. In.; 2017.
Illumina: Illumina Sequencing Technology. 2010.
Illumina: Infinium Assay workflow. 2012.
Inoue K, Lupski JR: Molecular mechanisms for genomic disorders. Annual review of genomics and human genetics 2002, 3(1):199-242.
Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature genetics 2012, 44(2):226.
Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavaré S: CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics 2010, 26(24):3051-3058.
Jeffreys AJ, Wilson V, Thein SL: Hypervariable'minisatellite' regions in human DNA. Nature 1985, 314(6006):67-73.
Jeffreys AJ, Wilson V, Thein SL: Individual-specific'fingerprints' of human DNA. Nature 1985, 316(6023):76-79.
Jenkins GM, Goddard ME, Black MA, Brauning R, Auvray B, Dodds KG, Kijas JW, Cockett N, McEwan JC: Copy number variants in the sheep genome detected using multiple approaches. BMC genomics 2016, 17(1):441.
Jia X, Chen S, Zhou H, Li D, Liu W, Yang N: Copy number variations identified in the chicken using a 60K SNP BeadChip. Animal genetics 2013, 44(3):276-284.
Jiang JC, Wang JY, Wang HF, Zhang Y, Kang HM, Feng XT, Wang JF, Yin ZJ, Bao WB, Zhang Q et al: Global copy number analyses by next generation sequencing provide insight into pig genome variation. BMC genomics 2014, 15.
Jiang L, Jiang J, Wang J, Ding X, Liu J, Zhang Q: Genome-wide identification of copy number variations in Chinese Holstein. PloS one 2012, 7(11):e48732.
Jiang L, Jiang JC, Yang J, Liu X, Wang JY, Wang HF, Ding XD, Liu JF, Zhang Q: Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. BMC genomics 2013, 14.
Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W et al: The sheep genome illuminates biology of the rumen and lipid metabolism. Science 2014, 344(6188):1168-1173.
Johnson AD, O'Donnell CJ: An Open Access Database of Genome-wide Association Results. BMC Medical Genetics 2009, 10:6-6.
Jung S, Yim S, Oh H, Park J, Kim M, Kim GA, Kim T, Kim J, Lee B, Chung Y: De novo copy number variations in cloned dogs from the same nuclear donor. BMC genomics 2013, 14(863).
Kader A, Liu X, Dong K, Song S, Pan J, Yang M, Chen X, He X, Jiang L, Ma Y: Identification of copy number variations in three Chinese horse breeds using 70K single nucleotide polymorphism BeadChip array. Animal genetics 2016, 47(5):560-569.
Kallioniemi A, Kallioniemi O-P, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992, 258(5083):818-821.
Kaushansky A, Albert SY, Austin LS, Mikolajczak SA, Vaughan AM, Camargo N, Metzger PG, Douglass AN, MacBeath G, Kappe SH: Suppression of host p53 is critical for Plasmodium liver-stage infection. Cell reports 2013, 3(3):630-637.

Kennedy SR, Loeb LA, Herr AJ: Somatic mutations in aging, cancer and neurodegeneration. Mechanisms of ageing and development 2012, 133(4):118-126.
Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F: Mapping and sequencing of structural variation from eight human genomes. Nature 2008, 453(7191):56-64.
Kijas J, Worley KC, Gibbs RA, Reid J, Yu F, Lee SL, Wu Y, Munzy DM, McWilliam S, Yu J et al: Whole genome sequencing of 75 sheep for variant detection and design of an HD chip. In: ISAG 33rd conference. Cairns, Australia; 2012.
Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S, Whan V: Analysis of copy number variants in the cattle genome. Gene 2011, 482(1):73-77.
Kijas JW: Haplotype-based analysis of selective sweeps in sheep. Genome 2014, 57(8):433-437.
Kim E-S, Elbeltagy A, Aboul-Naga A, Rischkowsky B, Sayre B, Mwacharo J, Rothschild M: Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. Heredity 2016, 116(3):255.
Kim J-H, Hu H-J, Yim S-H, Bae JS, Kim S-Y, Chung Y-J: CNVRuler: a copy number variation-based case–control association analysis tool. Bioinformatics 2012, 28(13):1790-1792.
Kim T-M, Luquette LJ, Xi R, Park PJ: rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. BMC bioinformatics 2010, 11(1):432.
Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, Hochreiter S: cn. MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic acids research 2012, 40(9):e69-e69.
Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al: Complement factor H polymorphism in age-related macular degeneration. Science 2005, 308(5720):385-389.
Kleinjan DA, van Heyningen V: Long-range control of gene expression: emerging mechanisms and disruption in disease. The American Journal of Human Genetics 2005, 76(1):8-32.
Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB: PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome biology 2009, 10(2):R23.
Krepischi AC, Achatz MIW, Santos EM, Costa SS, Lisboa BC, Brentani H, Santos TM, Gonçalves A, Nóbrega AF, Pearson PL: Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Research 2012, 14(1):R24.
Kurotaki N, Harada N, Shimokawa O, Miyake N, Kawame H, Uetake K, Makita Y, Kondoh T, Ogata T, Hasegawa T: Fifty microdeletions among 112 cases of Sotos syndrome: low copy repeats possibly mediate the common deletion. Human mutation 2003, 22(5):378-387.
LaFramboise T: Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic acids research 2009, 37(13):4181-4193.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W: Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.
Legault M-A, Girard S, Perreault L-PL, Rouleau GA, Dubé M-P: Comparison of sequencing based CNV discovery methods using monozygotic twin quartets. PloS one 2015, 10(3):e0122287.

Li H, Durbin R: Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009, 25.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The sequence alignment/map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.
Li W, Olivier M: Current analysis platforms and methods for detecting copy number variation. Physiological Genomics 2013, 45(1):1-16.
Li Y, Mei SQ, Zhang XY, Peng XW, Liu G, Tao H, Wu HY, Jiang SW, Xiong YZ, Li FG: Identification of genome-wide copy number variations among diverse pig breeds by array CGH. BMC genomics 2012, 13.
Litt M, Luty JA: A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. American journal of human genetics 1989, 44(3):397.
Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME: Analysis of copy number variations among diverse cattle breeds. Genome research 2010, 20(5):693-703.
Liu GE, Van Tassell CP, Sonstegard TS, Li RW, Alexander LJ, Keele JW, Matukumalli LK, Smith TP, Gasbarre LC: Detection of Germline and Somatic Copy Number Variations in Cattle. In: Animal Genomics for Animal Health. Edited by Pinard MH, Gay C, Pastoret PP, Dodet B, vol. 132. Basel: Karger; 2008: 231-237.
Liu J, Zhang L, Xu L, Ren H, Lu J, Zhang X, Zhang S, Zhou X, Wei C, Zhao F et al: Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. BMC genomics 2013, 14(229).
Liu Z, Ji Z, Wang G, Chao T, Hou L, Wang J: Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. BMC genomics 2016, 17(1):863.
Liu Z, Li A, Schulz V, Chen M, Tuck D: MixHMM: Inferring Copy Number Variation and Allelic Imbalance Using SNP Arrays and Tumor Samples Mixed with Stromal Cells. PLoS ONE 2010, 5(6):e10909.
Liu Z-W, Jarret RL, Duncan RR, Kresovich S: Genetic relationships and variation among ecotypes of seashore paspalum (Paspalum vaginatum) determined by random amplified polymorphic DNA markers. Genome 1994, 37(6):1011-1017.
Long X, Qiu X, Chen L, Wang J, Li X: Identification and analysis of copy number variations in rongchang pig hybridization population. Acta Veterinaria et Zootechnica Sinica 2014, 45(4):524-532.
Lowe A, Hanotte O, Guarino L: Standardization of molecular genetic techniques for the characterization of germplasm collections: the case of random amplified polymorphic DNA (RAPD). 1996.
Ma L, Chung WK: Quantitative analysis of copy number variants based on real-time LightCycler PCR. Curr Protoc Hum Genet 2014, 80:Unit 7 21.
Ma Q, Liu X, Pan J, Ma L, Ma Y, He X, Zhao Q, Pu Y, Li Y, Jiang L: Genome-wide detection of copy number variation in Chinese indigenous sheep using an ovine high-density 600 K SNP array. Scientific Reports 2017, 7(1):912.
Ma Y, Zhang Q, Lu Z, Zhao X, Zhang Y: Analysis of copy number variations by SNP50 BeadChip array in Chinese sheep. Genomics 2015, 106(5):295-300.
Ma Y, Zhang Q, Lu Z, Zhao X, Zhang Y: Analysis of copy number variations by SNP50 BeadChip array in Chinese sheep. Genomics 2015.
Magi A, Benelli M, Yoon S, Roviello F, Torricelli F: Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. Nucleic acids research 2011, 39(10):e65-e65.

Maheswaran M: Molecular markers: history features and applications. Advanced Biotech 2004, 51:373-378.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A: Finding the missing heritability of complex diseases. Nature 2009, 461(7265):747-753.
Mardis ER: The impact of next-generation sequencing technology on genetics. Trends in genetics 2008, 24(3):133-141.
Marshall K, Maddox J, Lee S, Zhang Y, Kahn L, Graser HU, Gondro C, Walkden-Brown S, Van Der Werf J: Genetic mapping of quantitative trait loci for resistance to Haemonchus contortus in sheep. Animal Genetics 2009, 40(3):262-272.
Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS: Development and characterization of a high density SNP genotyping assay for cattle. PloS one 2009, 4(4):e5350.
McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH: Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nature genetics 2008, 40(9):1107-1112.
McNally J, Murrell A: Detection of quantitative trait loci for internal parasite resistance in sheep. I. Linkage analysis in a Romney x Merino sheep backcross population. Parasitology 2010, 137:1275-1282.
McRae KM, McEwan JC, Dodds KG, Gemmell NJ: Signatures of selection in sheep bred for resistance or susceptibility to gastrointestinal nematodes. BMC genomics 2014, 15(1):1.
Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M: Detecting copy number variation with mated short reads. Genome research 2010, 20(11):1613-1622.
Metzger J, Philipp U, Lopes MS, Machado AD, Felicetti M, Silvestrelli M, Distl O: Analysis of copy number variants by three detection algorithms and their association with body size in horses. BMC genomics 2013, 14.
Michael KL, Taylor LC, Schultz SL, Walt DR: Randomly ordered addressable high-density optical sensor arrays. Analytical chemistry 1998, 70(7):1242-1248.
Miesfeld R, Krystal M, Amheim N: A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human δ and β globin genes. Nucleic acids research 1981, 9(22):5931-5948.
Miller CA, Hampton O, Coarfa C, Milosavljevic A: ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PloS one 2011, 6(1):e16327.
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK: Mapping copy number variation by population-scale genome sequencing. Nature 2011, 470(7332):59.
Molin AM, Berglund J, Webster MT, Lindblad-Toh K: Genome-wide copy number variant discovery in dogs using the CanineHD genotyping array. BMC genomics 2014, 15.
Morris C, Vlassoff A, Bisset S, Baker R, Watson T, West C, Wheeler M: Continued selection of Romney sheep for resistance or susceptibility to nematode infection: estimates of direct and correlated responses. Animal Science 2000, 70(1):17-27.
Morris C, Watson T, Bisset S, Vlassoff A, Douch P: Breeding sheep in New Zealand for resistance or resilience to nematode parasites. Breeding for resistance to infectious diseases in small ruminants 1995:77-98.
Morris C, Wheeler M, Watson T, Hosking B, Leathwick D: Direct and correlated responses to selection for high or low faecal nematode egg count in Perendale sheep. New Zealand Journal of Agricultural Research 2005, 48(1):1-10.
Morrow EM: Genomic copy number variation in disorders of cognitive development. Journal of the American Academy of Child & Adolescent Psychiatry 2010, 49(11):1091-1104.

Mueller JC, Andreoli C: Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity. Bioinformatics 2004, 20(5):786-787.
Nicholas TJ, Baker C, Eichler EE, Akey JM: A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. BMC genomics 2011, 12.
Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM: The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Research 2009, 19(3):491-499.
Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P: Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2002, 64(4):695-715.
Nijkamp JF, van den Broek MA, Geertman J-MA, Reinders MJ, Daran J-MG, de Ridder D: De novo detection of copy number variation by co-assembly. Bioinformatics 2012, 28(24):3195-3202.
Nishimura S, Watanabe T, Mizoshita K, Tatsuda K, Fujita T, Watanabe N, Sugimoto Y, Takasuga A: Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. BMC genetics 2012, 13(1):40.
Norris BJ, Whan VA: A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. Genome Research 2008, 18(8):1282-1293.
Nybom H: DNA fingerprinting—a useful tool in fruit breeding. In: Progress in Temperate Fruit Breeding. Springer; 1994: 257-262.
O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP: Extensive genetic variation in somatic human tissues. Proceedings of the National Academy of Sciences 2012, 109(44):18018-18023.
Oddy H, Dalrymple B, McEwan J, Kijas J, Haye B, Van Der Werf J, Emery D, Hynd P, Longhurst T, Fischer T: SheepGenomics and the International Sheep Genomics Consortium. In: Proceedings of the Association for the Advancement of Animal Breeding and Genetics: 2007; 2007: 411-417.
Oliphant A, Barker DL, Stuelpnagel JR, Chee MS: BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques 2002, 32(6):56-58.
Oluwole OA, Revay T, Mahboubi K, Favetta LA, King WA: Somatic mosaicism in bulls estimated from genome-wide CNV array and TSPY gene copy numbers. Cytogenetic and genome research 2016, 149(3):176-181.
Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui L-C: A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nature genetics 2001, 29(3):321-325.
Park KD, Kim H, Hwang JY, Lee CK, Do KT, Kim HS, Yang YM, Kwon YJ, Kim J, Kim HJ et al: Copy Number Deletion Has Little Impact on Gene Expression Levels in Racehorses. Asian-Australasian Journal of Animal Sciences 2014, 27(9):1345-1354.
Paterson S: Evidence for balancing selection at the major histocompatibility complex in a free-living ruminant. Journal of Heredity 1998, 89(4):289-294.
Paudel Y, Madsen O, Megens HJ, Frantz LAF, Bosse M, Bastiaansen JWM, Crooijmans R, Groenen MAM: Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC genomics 2013, 14.
Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J et al: High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 2006, 16(9):1136-1148.

Perry B, Randolph T: Improving the assessment of the economic impact of parasitic diseases and of their control in production animals. Veterinary parasitology 1999, 84(3):145-168.
Pickering NK: Genetics of flystrike, dagginess and associated traits in New Zealand dual-purpose sheep: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Animal Science at Massey University, Palmerston North, New Zealand. Massey University; 2013.
Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W-L, Chen C, Zhai Y: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature genetics 1998, 20(2):207-211.
Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R et al: Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 2011, 29.
Pique-Regi R, Cáceres A, González JR: R-Gada: a fast and flexible pipeline for copy number analysis in association studies. BMC bioinformatics 2010, 11:380-380.
Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S: Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 2008, 24(3):309-318.
Poduri A, Evrony GD, Cai X, Walsh CA: Somatic Mutation, Genomic Variation, and Neurological Disease. Science 2013, 341(6141).
Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A: The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 1996, 2(3):225-238.
Price AL, Zaitlen NA, Reich D, Patterson N: New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 2010, 11(7):459.
Primrose SB, Twyman R: Principles of gene manipulation and genomics: John Wiley & Sons; 2009.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ: PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 2007, 81(3):559-575.
Qi J, Zhao F: inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic acids research 2011, 39(suppl_2):W567-W575.
Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM: Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome research 2010, 20(5):623-635.
Ramayo-Caldas Y, Castelló A, Pena RN, Alves E, Mercadé A, Souza CA, Fernández AI, Perez-Enciso M, Folch JM: Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. BMC genomics 2010, 11(1):593.
Ramirez O, Olalde I, Berglund J, Lorente-Galdos B, Hernandez-Rodriguez J, Quilez J, Webster MT, Wayne RK, Lalueza-Fox C, Vila C et al: Analysis of structural diversity in wolf-like canids reveals post-domestication variants. BMC genomics 2014, 15.
Rao YS, Li J, Zhang R, Lin XR, Xu JG, Xie L, Xu ZQ, Wang L, Gan JK, Xie XJ et al: Copy number variation identification and analysis of the chicken genome using a 60K SNP BeadChip. Poultry Science 2016, 95(8):1750-1756.
Rattray P: Helminth parasites in the New Zealand Meat & Wool Pastoral Industries: A review of current issues. Final Report 2003:114-117.
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen WW et al: Global variation in copy number in the human genome. Nature 2006, 444(7118):444-454.
Reis-Filho JS: Next-generation sequencing. Breast Cancer Research 2009, 11(3):S12.

Reiter RS, Williams J, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA: Global and local genome mapping in Arabidopsis thaliana by using recombinant inbred lines and random amplified polymorphic DNAs. Proceedings of the National Academy of Sciences 1992, 89(4):1477-1481.
RESEARCHERS PUBLISH FULL NEANDERTHAL GENOME [http://ourworldsmysteries.com/researchers-publish-full-neanderthal-genome/]
Reymond A, Henrichsen CN, Harewood L, Merla G: Side effects of genome structural changes. Current opinion in genetics & development 2007, 17(5):381-386.
Rochus CM, Tortereau F, Plisson-Petit F, Restoux G, Moreno-Romieux C, Tosser-Klopp G, Servin B: High density genome scan for selection signatures in French sheep reveals allelic heterogeneity and introgression at adaptive loci. bioRxiv 2017:103010.
Rodríguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, Villa O, Hutchinson A, Earl J, Marenne G: Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. The American Journal of Human Genetics 2010, 87(1):129-138.
Roos MH: The role of drugs in the control of parasitic nematode infections: must we do without? Parasitology 1997, 114 Suppl:S137-144.
Rosengren Pielberg G, Golovko A, Sundstrom E, Curik I, Lennartsson J, Seltenhammer MH, Druml T, Binns M, Fitzsimmons C, Lindgren G et al: A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. Nature genetics 2008, 40(8):1004-1009.
Rubin C-J, Megens H-J, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB: Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences 2012, 109(48):19529-19536.
Rubin C-J, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S: Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 2010, 464(7288):587.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ: Detecting recent positive selection in the human genome from haplotype structure. Nature 2002, 419(6909):832-837.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al: Genome-wide detection and characterization of positive selection in human populations. Nature 2007, 449(7164):913-918.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001, 409(6822):928-933.
Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 2012, 485(7397):237-241.
Sasaki S, Watanabe T, Nishimura S, Sugimoto Y: Genome-wide identification of copy number variation using high-density single-nucleotide polymorphism array in Japanese Black cattle. BMC genetics 2016, 17(1):26.
Schaible RH: Developmental genetics of spotting patterns in the mouse. 1963.
Scheet P, Stephens M: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 2006, 78(4):629-644.
Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L: Challenges and standards in integrating surveys of structural variation. Nature genetics 2007, 39(7 Suppl):S7-15.

Schiavo G, Dolezal MA, Scotti E, Bertolini F, Calo DG, Galimberti G, Russo V, Fontanesi L: Copy number variants in Italian Large White pigs detected using high-density single nucleotide polymorphisms and their association with back fat thickness. Animal genetics 2014, 45(5):745-749.
Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M: Large-scale copy number polymorphism in the human genome. Science 2004, 305(5683):525-528.
Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, Ezra E, Zeron Y: Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. BMC genomics 2010, 11(1):673.
Shin DH, Lee HJ, Cho S, Kim HJ, Hwang JY, Lee CK, Jeong J, Yoon D, Kim H: Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level. BMC genomics 2014, 15.
Silva M, Sonstegard T, Hanotte O, Mugambi J, Garcia J, Nagda S, Gibson J, Iraqi F, McClintock A, Kemp S: Identification of quantitative trait loci affecting resistance to gastrointestinal parasites in a double backcross population of Red Maasai and Dorper sheep. Animal genetics 2012, 43(1):63-71.
Sindi S, Helman E, Bashir A, Raphael BJ: A geometric approach for classification and comparison of structural variants. Bioinformatics 2009, 25(12):i222-i230.
Sindi SS, Önal S, Peng LC, Wu H-T, Raphael BJ: An integrative probabilistic model for identification of structural variation in sequencing data. Genome biology 2012, 13(3):R22.
Skinner BM, Robertson LB, Tempest HG, Langley EJ, Ioannou D, Fowler KE, Crooijmans RP, Hall AD, Griffin DK, Völker M: Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis. BMC genomics 2009, 10(1):357.
Skinner DM, Beattie WG, Blattner FR, Stark BP, Dahlberg JE: Repeat sequence of a hermit crab satellite deoxyribonucleic acid is (-TAGG-) n.(-ATCC-) n. Biochemistry 1974, 13(19):3930-3937.
Slatkin M: Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 2008, 9(6):477-485.
Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, Cremer T, Lichter P: Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes, chromosomes and cancer 1997, 20(4):399-407.
Somavilla AL, Sonstegard T, Higa R, Rosa A, Siqueira F, Silva L, Torres Júnior R, Coutinho L, Mudadu M, Alencar M: A genome-wide scan for selection signatures in Nellore cattle. Animal genetics 2014, 45(6):771-781.
Spencer CCA, Su Z, Donnelly P, Marchini J: Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. PLOS Genetics 2009, 5(5):e1000477.
Sreenan JJ, Pettay JD, Tbakhi A, Totos G, Sandhaus LM, Miller ML, Bolwell B, Tubbs RR: The Use of Amplified Variable Number of Tandem Repeats (VNTR) in the Detection of Chimerism Following Bone Marrow Transplantation: A Comparison With Restriction Fragment Length Polymorphism (RFLP) by Southern Blotting. American Journal of Clinical Pathology 1997, 107(3):292-298.
Stankiewicz P, Lupski J: The genomic basis of disease, mechanisms and assays for genomic disorders. 2006.
Stella A, Ajmone-Marsan P, Lazzari B, Boettcher P: Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. Genetics 2010, 185(4):1451-1461.

Stothard P, Choi JW, Basu U, Sumner-Thomson JM, Meng Y, Liao XP, Moore SS: Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. BMC genomics 2011, 12.
Strachan T, Read A: Human Molecular Genetics (New York: Garland Science, Taylor & Francis Group). 2011.
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, De Grassi A, Lee C: Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007, 315(5813):848-853.
Strillacci MG, Cozzi MC, Gorla E, Mosca F, Schiavini F, Román-Ponce SI, Ruiz López FJ, Schiavone A, Marzoni M, Cerolini S et al: Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. animal 2016, 11(5):737-745.
Sun W, Wright FA, Tang Z, Nordgard SH, Loo PV, Yu T, Kristensen VN, Perou CM: Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic acids research 2009, 37(16):5365-5377.
Tang K, Thornton KR, Stoneking M: A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 2007, 5(7):e171.
Tanksley S, Young N, Paterson A, Bonierbale M: RFLP mapping in plant breeding: new tools for an old science. Nature Biotechnology 1989, 7(3):257-264.
Tautz D, Renz M: Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic acids research 1984, 12(10):4127-4138.
Tian M, Wang Y, Gu X, Feng C, Fang S, Hu X, Li N: Copy number variants in locally raised Chinese chicken genomes determined using array comparative genomic hybridization. BMC genomics 2013, 14(262).
Tingey SV, Rafalski JA, Hanafey MK: Genetic analysis with RAPD markers. In: Plant Molecular Biology. Springer; 1994: 491-500.
Tonegawa S: Somatic generation of antibody diversity. Nature 1983, 302(5909):575-581.
Turner SD: qqman: an R package for visualizing GWAS results using QQ and manhattan plots. BioRxiv 2014:005165.
Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D: Fine-scale structural variation of the human genome. Nature genetics 2005, 37(7):727-732.
Velasquez VLB, Gepts P: RFLP diversity of common bean (Phaseolus vulgaris) in its centres of origin. Genome 1994, 37(2):256-263.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA: The sequence of the human genome. science 2001, 291(5507):1304-1351.
Vitti JJ, Grossman SR, Sabeti PC: Detecting natural selection in genomic data. Annual review of genetics 2013, 47:97-120.
Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW: Cancer genome landscapes. science 2013, 339(6127):1546-1558.
Voight BF, Kudaravalli S, Wen X, Pritchard JK: A Map of Recent Positive Selection in the Human Genome. PLOS Biology 2006, 4(3):e72.
Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M: AFLP: a new technique for DNA fingerprinting. Nucleic acids research 1995, 23(21):4407-4414.
Voss H, Schwager C, Wiemann S, Zimmermann J, Stegemann J, Erfle H, Voie A-M, Drzonek H, Ansorge W: Efficient low redundancy large-scale DNA sequencing at EMBL. Journal of biotechnology 1995, 41(2):121-129.
Walt DR: Bead-based fiber-optic arrays. Science 2000, 287(5452):451-452.

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998, 280(5366):1077-1082.
Wang J, Jiang J, Fu W, Jiang L, Ding X, Liu J-F, Zhang Q: A genome-wide detection of copy number variations using SNP genotyping arrays in swine. BMC genomics 2012, 13(1):273.
Wang J, Wang H, Jiang J, Kang H, Feng X, Zhang Q, Liu J: Identification of genome-wide copy number variations among diverse pig breeds using SNP genotyping arrays. PLoS ONE 2013, 8(7).
Wang JY, Jiang JC, Wang HF, Kang HM, Zhang Q, Liu JF: Enhancing Genome-Wide Copy Number Variation Identification by High Density Array CGH Using Diverse Resources of Pig Breeds. Plos One 2014, 9(1).
Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M: PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome research 2007, 17(11):1665-1674.
Wang LG, Liu X, Zhang LC, Yan H, Luo WZ, Liang J, Cheng DX, Chen SK, Ma XJ, Song X et al: Genome-Wide Copy Number Variations Inferred from SNP Genotyping Arrays Using a Large White and Minzhu Intercross Population. Plos One 2013, 8(10).
Wang MD, Dzama K, Hefer CA, Muchadeyi FC: Genomic population structure and prevalence of copy number variations in South African Nguni cattle. BMC genomics 2015, 16(1):894.
Wang W, Wang SY, Hou CL, Xing YP, Cao JW, Wu KF, Liu CX, Zhang D, Zhang L, Zhang YR et al: Genome-Wide Detection of Copy Number Variations among Diverse Horse Breeds by Array CGH. Plos One 2014, 9(1).
Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N: An initial map of chromosomal segmental copy number variations in the chicken. BMC genomics 2010, 11(1):351.
Wang Y, Gu X, Feng C, Song C, Hu X, Li N: A genome-wide survey of copy number variation regions in various chicken breeds by array comparative genomic hybridization method. Animal Genetics 2012, 43(3):282-289.
Wang Y, Li J, Kolon TF, Fisher AO, Figueroa TE, BaniHani AH, Hagerty JA, Gonzalez R, Noh PH, Chiavacci RM: Genomic copy number variation association study in Caucasian patients with nonsyndromic cryptorchidism. BMC urology 2016, 16(1):62.
Wang Y, Tang Z, Sun Y, Wang H, Wang C, Yu S, Liu J, Zhang Y, Fan B, Li K et al: Analysis of genome-wide copy number variations in chinese indigenous and Western pig breeds by 60 k SNP genotyping arrays. PloS one 2014, 9(9):e106780-e106780.
Wang Z, Chen Q, Liao R, Zhang Z, Zhang X, Liu X, Zhu M, Zhang W, Xue M, Yang H: Genome-wide genetic variation discovery in Chinese Taihu pig breeds using next generation sequencing. Animal genetics 2017, 48(1):38-47.
Wang Z, Hormozdiari F, Yang W-Y, Halperin E, Eskin E: CNVeM: copy number variation detection using uncertainty of read mapping. Journal of Computational Biology 2013, 20(3):224-236.
Wang Z, Weber JL, Zhong G, Tanksley S: Survey of plant short tandem DNA repeats. Theoretical and applied genetics 1994, 88(1):1-6.
Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. evolution 1984, 38(6):1358-1370.
Weir BS, Hill WG: Estimating F-statistics. Annual review of genetics 2002, 36(1):721-750.
Welsh J, McClelland M: Fingerprinting genomes using PCR with arbitrary primers. Nucleic acids research 1990, 18(24):7213-7218.

What is a copy number variant, and why are they important risk factors for ASD [http://readingroom.mindspec.org/?page_id=8221]
Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic acids research 1990, 18(22):6531-6535.
Winchester L, Yau C, Ragoussis J: Comparing CNV detection methods for SNP arrays. Briefings in functional genomics & proteomics 2009, 8(5):353-366.
Wright D, Boije H, Meadows JR, Bed'Hom B, Gourichon D, Vieaud A, Tixier-Boichard M, Rubin C-J, Imsland F, Hallböök F: Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS genetics 2009, 5(6):e1000512.
Wu Y, Fan H, Jing S, Xia J, Chen Y, Zhang L, Gao X, Li J, Gao H, Ren H: A genome-wide scan for copy number variations using high-density single nucleotide polymorphism array in Simmental cattle. Animal genetics 2015, 46(3):289-298.
Xi R, Hadjipanayis AG, Luquette LJ, Kim T-M, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA: Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proceedings of the National Academy of Sciences 2011, 108(46):E1128-E1136.
Xie C, Tammi MT: CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC bioinformatics 2009, 10(1):80.
Xu L, Hou Y, Bickhart D, Song J, Liu G: Comparative Analysis of CNV Calling Algorithms: Literature Survey and a Case Study Using Bovine High-Density SNP Data. Microarrays 2013, 2(3):171-185.
Xu L, Hou Y, Bickhart DM, Zhou Y, Song J, Sonstegard TS, Van Tassell CP, Liu GE: Population-genetic properties of differentiated copy number variations in cattle. Scientific reports 2016, 6:23161.
Yan J, Blair HT, Liu M, Li W, He S, Chen L, Dittmer KE, Garrick DJ, Biggs PJ, Dukkipati VS: Genome-wide detection of autosomal copy number variants in several sheep breeds using Illumina OvineSNP50 BeadChips. Small Ruminant Research 2017, 155:24-32.
Yan J, Blair HT, Liu M, Li W, He S, Chen L, Dittmer KE, Garrick DJ, Biggs PJ, Dukkipati VSR: Genome-wide detection of autosomal copy number variants in several sheep breeds using Illumina OvineSNP50 BeadChips. Small Ruminant Research 2017, 155(Supplement C):24-32.
Yan Y: Genome-Wide Assessment of Copy Number Variation in Two Chicken Lines with Different Susceptibility to Marek's Disease Using Next Generation Sequencing. In: Plant and Animal Genome XXIII Conference: 2015: Plant and Animal Genome; 2015.
Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K: Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. The American Journal of Human Genetics 2007, 80(6):1037-1054.
Yang L, Xu L, Zhou Y, Liu M, Wang L, Kijas JW, Zhang H, Li L, Liu GE: Diversity of copy number variation in a worldwide population of sheep. Genomics 2018, 110(3):143-148.
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009, 25(21):2865-2871.
Yi GQ, Qu LJ, Liu JF, Yan YY, Xu GY, Yang N: Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC genomics 2014, 15.

Yindom LM, Forbes R, Aka P, Janha O, Jeffries D, Jallow M, Conway DJ, Walther M: Killer-cell immunoglobulin-like receptors and malaria caused by Plasmodium falciparum in The Gambia. Tissue Antigens 2012, 79(2):104-113.
Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 2009, 19(9):1586-1592.
Yuan Z, Liu E, Liu Z, Kijas J, Zhu C, Hu S, Ma X, Zhang L, Du L, Wang H: Selection signature analysis reveals genes associated with tail type in Chinese indigenous sheep. Animal genetics 2017, 48(1):55-66.
Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-Né P, Nicolas A, Delattre O, Barillot E: SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 2010, 26(15):1895-1896.
Zerr T, Cooper GM, Eichler EE, Nickerson DA: Targeted interrogation of copy number variation using SCIMMkit. Bioinformatics 2010, 26(1):120-122.
Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C: Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC genomics 2011, 12(1):557.
Zhang H, Du ZQ, Dong JQ, Wang HX, Shi HY, Wang N, Wang SZ, Li H: Detection of genome-wide copy number variations in two chicken lines divergently selected for abdominal fat content. BMC genomics 2014, 15.
Zhang H, Wang S-Z, Wang Z-P, Da Y, Wang N, Hu X-X, Zhang Y-D, Wang Y-X, Leng L, Tang Z-Q: A genome-wide scan of selective sweeps in two broiler chicken lines divergently selected for abdominal fat content. BMC genomics 2012, 13(1):1.
Zhang J, Wu Y: SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data. Bioinformatics 2011, 27(23):3228-3234.
Zhang L, Liu J, Zhao F, Ren H, Xu L, Lu J, Zhang S, Zhang X, Wei C, Lu G et al: Genome-Wide Association Studies for Growth and Meat Production Traits in Sheep. PLOS ONE 2013, 8(6):e66569.
Zhang LZ, Jia SG, Yang MJ, Xu Y, Li CJ, Sun JJ, Huang YZ, Lan XY, Lei CZ, Zhou Y et al: Detection of copy number variations and their effects in Chinese bulls. BMC genomics 2014, 15.
Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, Chen K, Shi X, Kraja A, Mardis ER, Wilson RK: CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. Bioinformatics 2010, 26(4):464-469.
Zhang Q, Ma Y, Wang X, Zhang Y, Zhao X: Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. Molecular genetics and genomics : MGG 2015, 290(1).
Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, Gerstein M: Identification of genomic indels and structural variations using split reads. BMC genomics 2011, 12(1):375.
Zhao F-p, Wei C-h, Zhang L, Liu J-s, Wang G-k, Zeng T, Du L-x: A genome scan of recent positive selection signatures in three sheep populations. Journal of Integrative Agriculture 2016, 15(1):162-174.
Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC bioinformatics 2013, 14(11):S1.
Zhao X, Dittmer KE, Blair HT, Thompson KG, Rothschild MF, Garrick DJ: A Novel Nonsense Mutation in the DMP1 Gene Identified by a Genome-Wide Association Study Is Responsible for Inherited Rickets in Corriedale Sheep. PLoS ONE 2011, 6(7):e21739.

Zhao X, Onteru SK, Dittmer KE, Parton K, Blair HT, Rothschild MF, Garrick DJ: A missense mutation in AGTPBP1 was identified in sheep with a lower motor neuron disease. Heredity 2012, 109(3):156-162.
Zhi D, Da L, Liu M, Cheng C, Zhang Y, Wang X, Li X, Tian Z, Yang Y, He T: Whole Genome Sequencing of Hulunbuir Short-Tailed Sheep for Identifying Candidate Genes Related to the Short-Tail Phenotype. G3: Genes, Genomes, Genetics 2018, 8(2):377-383.
Zhou L, Li J, Yang J, Liu C, Xie X, He Y, Liu X, Xin W, Zhang W, Ren J: Genome-wide mapping of copy number variations in commercial hybrid pigs using a high-density SNP genotyping array. Russian journal of genetics 2016, 52(1):85-92.
Zhou W, Liu RR, Zhang JJ, Zheng MQ, Li P, Chang GB, Wen J, Zhao GP: A genome-wide detection of copy number variation using SNP genotyping arrays in Beijing-You chickens. Genetica 2014, 142(5):441-450.
Zhu C, Fan H, Yuan Z, Hu S, Ma X, Xuan J, Wang H, Zhang L, Wei C, Zhang Q: Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays. Scientific reports 2016, 6:27822.
Zhu C, Fan H, Yuan Z, Hu S, Zhang L, Wei C, Zhang Q, Zhao F, Du L: Detection of selection signatures on the X chromosome in three sheep breeds. International journal of molecular sciences 2015, 16(9):20360-20374.
Žilina O, Koltšina M, Raid R, Kurg A, Tõnisson N, Salumets A: Somatic mosaicism for copy-neutral loss of heterozygosity and DNA copy number variations in the human genome. BMC genomics 2015, 16(1):703.

Appendix

3.1 Code for NGS data mapping

```
# Uncompress file
gunzip H2KJYBCXX-828-05-1_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-1_TGACCA_L002_R1_001.fastq.gz
gunzip H2L55BCXX-828-05-1_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-1_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-1_TGACCA_L002_R2_001.fastq.gz
gunzip H2L55BCXX-828-05-1_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-2_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-2_TGACCA_L002_R1_001.fastq.gz
gunzip H2L55BCXX-828-05-2_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-2_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-2_TGACCA_L002_R2_001.fastq.gz
gunzip H2L55BCXX-828-05-2_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-3_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-3_TGACCA_L002_R1_001.fastq.gz
gunzip H2L55BCXX-828-05-3_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-3_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-3_TGACCA_L002_R2_001.fastq.gz
gunzip H2L55BCXX-828-05-3_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-4_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-4_TGACCA_L002_R1_001.fastq.gz
gunzip H2L55BCXX-828-05-4_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-4_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-4_TGACCA_L002_R2_001.fastq.gz
gunzip H2L55BCXX-828-05-4_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-5_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-5_TGACCA_L002_R1_001.fastq.gz
gunzip H2L55BCXX-828-05-5_TGACCA_L001_R1_001.fastq.gz
gunzip H2KJYBCXX-828-05-5_TGACCA_L001_R2_001.fastq.gz
gunzip H2KJYBCXX-828-05-5_TGACCA_L002_R2_001.fastq.gz
gunzip H2L55BCXX-828-05-5_TGACCA_L001_R2_001.fastq.gz

# Merge two reads
cat       H2KJYBCXX-828-05-1_TGACCA_L001_R1_001.fastq       H2KJYBCXX-828-05-
1_TGACCA_L002_R1_001.fastq H2L55BCXX-828-05-1_TGACCA_L001_R1_001.fastq > 828-
05-1_R1.fastq
cat       H2KJYBCXX-828-05-1_TGACCA_L001_R2_001.fastq       H2KJYBCXX-828-05-
1_TGACCA_L002_R2_001.fastq H2L55BCXX-828-05-1_TGACCA_L001_R2_001.fastq > 828-
05-1_R2.fastq
cat       H2KJYBCXX-828-05-2_TGACCA_L001_R1_001.fastq       H2KJYBCXX-828-05-
2_TGACCA_L002_R1_001.fastq H2L55BCXX-828-05-2_TGACCA_L001_R1_001.fastq > 828-
05-2_R1.fastq
cat       H2KJYBCXX-828-05-2_TGACCA_L001_R2_001.fastq       H2KJYBCXX-828-05-
2_TGACCA_L002_R2_001.fastq H2L55BCXX-828-05-2_TGACCA_L001_R2_001.fastq > 828-
05-2_R2.fastq
cat       H2KJYBCXX-828-05-3_TGACCA_L001_R1_001.fastq       H2KJYBCXX-828-05-
3_TGACCA_L002_R1_001.fastq H2L55BCXX-828-05-3_TGACCA_L001_R1_001.fastq > 828-
05-3_R1.fastq
cat       H2KJYBCXX-828-05-3_TGACCA_L001_R2_001.fastq       H2KJYBCXX-828-05-
3_TGACCA_L002_R2_001.fastq H2L55BCXX-828-05-3_TGACCA_L001_R2_001.fastq > 828-
05-3_R2.fastq
cat       H2KJYBCXX-828-05-4_TGACCA_L001_R1_001.fastq       H2KJYBCXX-828-05-
```

```
4_TGACCA_L002_R1_001.fastq H2L55BCXX-828-05-4_TGACCA_L001_R1_001.fastq > 828-
05-4_R1.fastq
cat       H2KJYBCXX-828-05-4_TGACCA_L001_R2_001.fastq       H2KJYBCXX-828-05-
4_TGACCA_L002_R2_001.fastq H2L55BCXX-828-05-4_TGACCA_L001_R2_001.fastq > 828-
05-4_R2.fastq
cat       H2KJYBCXX-828-05-5_TGACCA_L001_R1_001.fastq       H2KJYBCXX-828-05-
5_TGACCA_L002_R1_001.fastq H2L55BCXX-828-05-5_TGACCA_L001_R1_001.fastq > 828-
05-5_R1.fastq
cat       H2KJYBCXX-828-05-5_TGACCA_L001_R2_001.fastq       H2KJYBCXX-828-05-
5_TGACCA_L002_R2_001.fastq H2L55BCXX-828-05-5_TGACCA_L001_R2_001.fastq > 828-
05-5_R2.fastq

# Create index for BWA
bwa index reference.fa
# Map reads to reference
bwa mem -M -R '@RG\tID:2\tSM:828-05-1\tPL:illumina\tLB:lib1' reference.fa 828-
05-1_R1.fastq 828-05-1_R2.fastq > 828-05-1_aligned.sam
bwa mem -M -R '@RG\tID:2\tSM:828-05-2\tPL:illumina\tLB:lib1' reference.fa 828-
05-2_R1.fastq 828-05-2_R2.fastq > 828-05-2_aligned.sam
bwa mem -M -R '@RG\tID:2\tSM:828-05-3\tPL:illumina\tLB:lib1' reference.fa 828-
05-3_R1.fastq 828-05-3_R2.fastq > 828-05-3_aligned.sam
bwa mem -M -R '@RG\tID:2\tSM:828-05-4\tPL:illumina\tLB:lib1' reference.fa 828-
05-4_R1.fastq 828-05-4_R2.fastq > 828-05-4_aligned.sam
bwa mem -M -R '@RG\tID:2\tSM:828-05-5\tPL:illumina\tLB:lib1' reference.fa 828-
05-5_R1.fastq 828-05-5_R2.fastq > 828-05-5_aligned.sam

 # Convert SAM to BAM
samtools view -@ 8 –b –S 828-05-1_aligned.sam > 828-05-1_aligned.bam
samtools view -@ 8 –b –S 828-05-2_aligned.sam > 828-05-2_aligned.bam
samtools view -@ 8 –b –S 828-05-3_aligned.sam > 828-05-3_aligned.bam
samtools view -@ 8 –b –S 828-05-4_aligned.sam > 828-05-4_aligned.bam
samtools view -@ 8 –b –S 828-05-5_aligned.sam > 828-05-5_aligned.bam

 # Sort BAM
samtools sort -@ 8 –o 828-05-2_sorted.bam 828-05-1_aligned.bam
samtools sort -@ 8 –o 828-05-2_sorted.bam 828-05-2_aligned.bam
samtools sort -@ 8 –o 828-05-2_sorted.bam 828-05-3_aligned.bam
samtools sort -@ 8 –o 828-05-2_sorted.bam 828-05-4_aligned.bam
samtools sort -@ 8 –o 828-05-2_sorted.bam 828-05-5_aligned.bam

# Duplicates Marking
java   -jar   /usr/share/java/picard/picard.jar   MarkDuplicates   INPUT=828-05-
1_LL_sorted.bam   OUTPUT=828-05-1_LL_markdup.bam   METRICS_FILE=metrics_828-05-
1.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1020
java   -jar   /usr/share/java/picard/picard.jar   MarkDuplicates   INPUT=828-05-
2_LL_sorted.bam   OUTPUT=828-05-2_LL_markdup.bam   METRICS_FILE=metrics_828-05-
2.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1020
java   -jar   /usr/share/java/picard/picard.jar   MarkDuplicates   INPUT=828-05-
3_LL_sorted.bam   OUTPUT=828-05-3_LL_markdup.bam   METRICS_FILE=metrics_828-05-
3.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1020
java   -jar   /usr/share/java/picard/picard.jar   MarkDuplicates   INPUT=828-05-
4_LL_sorted.bam   OUTPUT=828-05-4_LL_markdup.bam   METRICS_FILE=metrics_828-05-
4.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1020
java   -jar   /usr/share/java/picard/picard.jar   MarkDuplicates   INPUT=828-05-
5_LL_sorted.bam   OUTPUT=828-05-5_LL_markdup.bam   METRICS_FILE=metrics_828-05-
5.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1020
```

3.2 R code for CNVR plot

```
chromosome <- c(275612895,248993846,224283230,119255633,107901688,117031472,

100079507,90695168,94726778,86447213,62248096,79100223,83079144,

62722625,80923592,71719816,72286588,68604602,60464314,51176841,

50073674,50832532,62330649,42034648,45367442,44077779)
rnames <- c("CNVR")
cnames <- c("Chr1","Chr2","Chr3","Chr4","Chr5","Chr6",
                    "Chr7","Chr8","Chr9","Chr10","Chr11","Chr12","Chr13",
                    "Chr14","Chr15","Chr16","Chr17","Chr18","Chr19","Chr20",
                    "Chr21","Chr22","Chr23","Chr24","Chr25","Chr26")
opar <- par(no.readonly=TRUE)
barplot(chromosome,  beside=T,  horiz=T,  xlab="Chromosome  Length  (Mbp)",
xlim=c(0,300000000), axes=FALSE, cex.names=0.8, las=1, col=0, names.arg=cnames)
axis(1,
at=c(0,30000000,60000000,90000000,120000000,150000000,
180000000,210000000,240000000,270000000,300000000),
labels=c(0,30,60,90,120,150,180,210,240,270,300), las=1)
legend(locator(1),     title="Event",     c("Loss","Gain","Mix"),     pch=22,
pt.bg=c("blue","red","yellow"))
attach(loss)
segments(Start, (Chr-1)*1.2+0.7, End, (Chr-1)*1.2+0.7,col="blue",lty=1,lwd=4)
detach(loss)
attach(gain)
segments(Start, (Chr-1)*1.2 + 0.7 ,End ,(Chr-1)*1.2+0.7,col="red",lty=1,lwd=4)
detach(gain)
attach(mix)
segments(Start, (Chr-1)*1.2 +0.7 ,End, (Chr-1)*1.2+0.7,col="yellow",lty=1,lwd=4)
detach(mix)
par(opar)
```

3.3 Code for exploring sequencing depth in five samples

```
install.packages("devtools")
library(devtools)
install_github("easyGgplot2", "kassambara")l
library(easyGgplot2)
library(ggplot2)
library(reshape2)
library(gridExtra)
all_sample <- read.table('all_chr13_50kb')
CNV_1 <- read.table('CNV_1', header = TRUE)
CNV_2 <- read.table('CNV_2', header = TRUE)
CNV_3 <- read.table('CNV_3', header = TRUE)
CNV_4 <- read.table('CNV_4', header = TRUE)
CNV_5 <- read.table('CNV_5', header = TRUE)
names(all_sample)   <-   c('position','sample_1',   'sample_2',   'sample_3',
'sample_4', 'sample_5')
plot_a  <-ggplot(all_sample,  aes(x=position,  y=sample_1))  +  geom_line()  +
theme_bw() +
 xlab("Chromosome 13") + ylab("Depth") + ggtitle("828-05-01") + ylim(5,15) +
 geom_point(data=CNV_1, aes(x=position, y=value))
plot_b <-ggplot(all_sample, aes(x=position, y=sample_2)) + geom_line() +
 theme_bw() + xlab("Chromosome 13") + ylab("Depth") + ggtitle("828-05-02") +
ylim(5,15) +
 geom_point(data=CNV_2, aes(x=position, y=value))
plot_c <-ggplot(all_sample, aes(x=position, y=sample_3)) + geom_line() +
 theme_bw() + xlab("Chromosome 13") + ylab("Depth") + ggtitle("828-05-03") +
ylim(5,15) +
  geom_point(data=CNV_3, aes(x=position, y=value))
plot_d <-ggplot(all_sample, aes(x=position, y=sample_4)) + geom_line() +
 theme_bw() + xlab("Chromosome 13") + ylab("Depth") + ggtitle("828-05-04") +
ylim(5,15) +
 geom_point(data=CNV_4, aes(x=position, y=value))
plot_e  <-ggplot(all_sample,  aes(x=position,  y=sample_5))  +  geom_line()  +
theme_bw() +
 xlab("Chromosome 13") + ylab("Depth") + ggtitle("828-05-05") + ylim(5,15) +
 geom_point(data=CNV_5, aes(x=position, y=value))
grid.arrange(plot_a, plot_b, plot_c, plot_d, plot_e, ncol=2)
```

3.4 Code for 10kb bin

```
with open('828-05-1_bin.depth','w') as output_file:
    with open('828-05-1_10kb.depth') as input_file:
        for line in input_file:
            if float(line.split()[0]) <=5:
                output_file.write('5'+'\n')
            elif float(line.split()[0]) > 5 and float(line.split()[0]) <= 100:
                output_file.write(str(round(float(line.split()[0])))+'\n')
            elif float(line.split()[0]) > 100:
                output_file.write('100' +'\n')

with open('828-05-2_bin.depth','w') as output_file:
    with open('828-05-2_10kb.depth') as input_file:
        for line in input_file:
            if float(line.split()[0]) <=5:
                output_file.write('5'+'\n')
            elif float(line.split()[0]) > 5 and float(line.split()[0]) <= 100:
                output_file.write(str(round(float(line.split()[0])))+'\n')
            elif float(line.split()[0]) > 100:
                output_file.write('100' +'\n')

with open('828-05-3_bin.depth','w') as output_file:
    with open('828-05-3_10kb.depth') as input_file:
        for line in input_file:
            if float(line.split()[0]) <=5:
                output_file.write('5'+'\n')
            elif float(line.split()[0]) > 5 and float(line.split()[0]) <= 100:
                output_file.write(str(round(float(line.split()[0])))+'\n')
            elif float(line.split()[0]) > 100:
                output_file.write('100' +'\n')

with open('828-05-4_bin.depth','w') as output_file:
    with open('828-05-4_10kb.depth') as input_file:
        for line in input_file:
            if float(line.split()[0]) <=5:
                output_file.write('5'+'\n')
            elif float(line.split()[0]) > 5 and float(line.split()[0]) <= 100:
                output_file.write(str(round(float(line.split()[0])))+'\n')
            elif float(line.split()[0]) > 100:
                output_file.write('100' +'\n')

with open('828-05-5_bin.depth','w') as output_file:
    with open('828-05-5_10kb.depth') as input_file:
        for line in input_file:
            if float(line.split()[0]) <=5:
                output_file.write('5'+'\n')
            elif float(line.split()[0]) > 5 and float(line.split()[0]) <= 100:
                output_file.write(str(round(float(line.split()[0])))+'\n')
            elif float(line.split()[0]) > 100:
                output_file.write('100' +'\n')
```

3.5 Code for violin plot

```
library(vioplot)
library(sm)
x1 <-read.table('828-05-1_10kb.depth')$V1
x2 <-read.table('828-05-2_10kb.depth')$V1
x3 <-read.table('828-05-3_10kb.depth')$V1
x4 <-read.table('828-05-4_10kb.depth')$V1
x5 <-read.table('828-05-5_10kb.depth')$V1

vioplot(log10(x1), log10(x2), log10(x3), log10(x4), log10(x5),
       names=c("828-05-1", "828-05-2", "828-05-3", "828-05-4", "828-05-5"),
       col="gold")
title("Violin Plots of Depth")
mtext("log10(Average Depth) of 10Kb",side=2,line=2.5,cex=1,font=2)
mtext("Sample ID",side=1,line=2.5,cex=1,font=2)
```

4.1 Custom written script in Perl and SQL for gene annotation

```
#!/usr/bin/perl
#
#   created by:     pjb on  2017-02-05
#   last edited by: pjb on  2017-02-05
#
#   A script to analyse the 600k SNP sheep dataset to find regions of
#   varying sizes where SNPs are found to coincide.  Output can be numbers
#   of SNPs per bin per statistical test, as well as the actual SNPs.
#
#   Data can be generated under a variety of combinations of negative
#   log values and bin sizes as definde by the user.  Bin sizes are in kb.
#
#   The script needs to be run initially to load in the data ("-rerun no"),
#   but after that can be run with any combination of values and bin sizes.
#
#########################################################

use warnings;
use strict;
use DBI;
use Bio::SeqIO;
use Bio::Seq;
use Getopt::Long;
use POSIX;

my ($binSize, $logValue, $rerun);

GetOptions( 'binSize:s'     => \$binSize,
            'logValue:s'    => \$logValue,
            'rerun:s'       => \$rerun);

my  ($statement,  $joiner,  $dbh,  $sth,  $datasource,  $querystring,  $rowcount,
$count);
my ($curT, $empty, $inFile, $inTable, $value, $negLogVal, $i, $j, $outFile,
$valTable);

my @tables  = ('S2_ihs_resistant', 'S3_ihs_resilient', 'S4_xpehh', 'S5_rsb');


## global variables ##

my $base        = ("/media/sf_macDocuments/Massey/2017/sheepSNPs/");
my $results     = ($base . "results/");
my $overallT1   = ("johnSNPoverallStart");
my $overallT    = ("johnSNPoverall");
my $selectedT1  = ("johnSNPselectedStart");
my $selectedT   = ("johnSNPselected");
my $chrsSum     = ($results . "chromosomeStat.txt");

my $compact     = ("bin_" . $binSize . "_negLog" . $logValue);
my $localRes    = ($results . $compact . "/");

if (-e $localRes) { print ("$localRes already exists.\n");
} else {            system "mkdir $localRes";
```

```
}


## on with the work ##

my $log     = ($localRes . "processLog.txt");

open (LOG, ">$log") or die ("couldn't open $log: $!\n");

print ("The process was started at " . scalar(localtime) . ".\n");
print LOG ("The process was started at " . scalar(localtime) . ".\n");


## set up db connection ##

&dbConnect();


## perform work as appropriate ##

if ($rerun eq 'yes') {

    print ("We have already done the large amount of database work.\n");
    print LOG ("We have already done the large amount of database work.\n");

} elsif ($rerun eq 'no') {

    ## parse the SNPs ##

    foreach $curT (@tables) {
        $inFile     = ($base . "source/Table" . $curT . "All.txt");
        $inTable    = ("Table" . $curT);

        &SNPstart($curT, $inTable, $inFile);
    }


    ## join data and find chromosomal ends ##

    &dataAnalysis();
}


## generation of outputs ##

&dataBinning($binSize, $logValue);


print ("\nProcess finished at " . scalar(localtime) . ".\n");
print LOG ("\nProcess finished at " . scalar(localtime) . ".\n");

close LOG;


####################
#                  #
```

```
#   subroutines     #
#                   #
#####################

sub SNPstart {
    ($curT, $inTable, $inFile)  = @_;

    $sth      =      $dbh->prepare      (qq{drop      table      if      exists
$inTable});    $sth->execute();
    $sth  =  $dbh->prepare  (qq{create  table  $inTable  (SNPname  varchar(30),
chromosome   smallint,   position   int,   value   float,   negLogVal   float)});
$sth->execute();
    $sth    =    $dbh->prepare    (qq{alter    table    $inTable    add    index
index1(SNPname)});   $sth->execute();

    $sth  =  $dbh->prepare  (qq{load  data  local  infile  '$inFile'  into  table
$inTable ignore 1 lines});  $sth->execute();

    print ("Table $curT created at " . scalar(localtime) . ".\n");
    print LOG ("Table $curT created at " . scalar(localtime) . ".\n");

    return ($curT, $inTable, $inFile);
}


sub dataAnalysis {

    my $T2      = ("TableS2_ihs_resistant");
    my $T3      = ("TableS3_ihs_resilient");
    my $T4      = ("TableS4_xpehh");
    my $T5      = ("TableS5_rsb");


    ## create overall table ##

    $sth = $dbh->prepare (qq{drop table if exists $overallT});   $sth->execute();
    $sth = $dbh->prepare (qq{drop table if exists $overallT1});  $sth->execute();
    $sth = $dbh->prepare (qq{create table $overallT1 select SNPname, chromosome,
position  from  $T2  union  select  SNPname,  chromosome,  position  from  $T3  union
select SNPname, chromosome, position from $T4 union select SNPname, chromosome,
position from $T5});    $sth->execute();
    $sth = $dbh->prepare (qq{alter table $overallT1 add index index1(SNPname)});
$sth->execute();

    $sth  =  $dbh->prepare  (qq{create  table  $overallT  select  *  from  $overallT1
order by chromosome, position});    $sth->execute();
    $sth    =    $dbh->prepare    (qq{alter    table    $overallT    add    index
index1(SNPname)});  $sth->execute();

    $sth = $dbh->prepare (qq{drop table if exists $overallT1});  $sth->execute();


    ## add in all SNP values ##

    foreach $curT (@tables) {
        $inTable    = ("Table" . $curT);
```

```
        $value      = ("val" . $curT);
        $negLogVal  = ("negLogVal" . $curT);

        $sth = $dbh->prepare (qq{alter table $overallT add column $value
float});    $sth->execute();
        $sth = $dbh->prepare (qq{alter table $overallT add column $negLogVal
float});    $sth->execute();

        $sth = $dbh->prepare (qq{update $overallT o, $inTable t set o.$value =
t.value where o.SNPname = t.SNPname});    $sth->execute();
        $sth = $dbh->prepare (qq{update $overallT o, $inTable t set o.$negLogVal
= t.negLogVal where o.SNPname = t.SNPname});    $sth->execute();

        print ("Table $overallT updated with data from $curT at " .
scalar(localtime) . ".\n");
        print LOG ("Table $overallT updated with data from $curT at " .
scalar(localtime) . ".\n");
    }


    ## generate max SNP location per chromosome ##

    open (OUT, ">$chrsSum") or die ("couldn't open $chrsSum: $!\n");

    print OUT ("chromosome\tmaxSNP\tnumberSNPs\n");

    $statement = ("select chromosome, max(position) as maxSNP, count(*) as
numberSNPs from $overallT group by chromosome");
    &statementPull ($statement, "\t");

    close OUT;
}


sub dataBinning {
    ($binSize, $logValue)   = @_;

    my $realBin     = ($binSize * 1000);
    my $SNPres      = ($localRes . "binSNPsOverall.txt");
    my $binFile     = ($localRes . "binTable.txt");
    my $sortedBin   = ($localRes . "orderedOutput.txt");
    my $binTable    = ("binFor_" . $realBin);
    my $localBins   = ($localRes . "tempThing.txt");

    ## create the bins required into a new table ##

    open (OUT, ">$binFile") or die ("couldn't open $binFile: $!\n");

    print OUT ("chromosome\tbinStart\tbinEnd\n");

    for ($i = 1; $i <= 26; $i++) {

        open (IN1, "<$chrsSum") or die ("couldn't open $chrsSum: $!\n");

        while (<IN1>) {
            chomp;
```

```
            my ($lChrs, $lMax, $lNum)   = split;

            if ($lChrs eq $i) {
                print ("We are working with chromosome $i at " .
scalar(localtime) . ".\n");
                print LOG ("We are working with chromosome $i at " .
scalar(localtime) . ".\n");

                my $maxBinNum   = ceil($lMax/$realBin);

                print ("\tFor chromosome $i with a binsize of $realBin bp, we
have $maxBinNum bins to analyse...\n");
                print LOG ("\tFor chromosome $i with a binsize of $realBin bp,
we have $maxBinNum bins to analyse...\n");

                for ($j = 1; $j <= $maxBinNum; $j++) {
                    my $start   = (($j - 1) * $realBin) + 1;
                    my $end     = ($j * $realBin);

                    print OUT ("$lChrs\t$start\t$end\n");
                }
            }
        }

        close IN1;

    }

    close OUT;

    $sth = $dbh->prepare (qq{drop table if exists $binTable});   $sth->execute();
    $sth = $dbh->prepare (qq{create table $binTable (chromosome smallint,
binStart int, binEnd int)});   $sth->execute();
    $sth    =    $dbh->prepare    (qq{alter    table    $binTable    add    index
index1(chromosome)});   $sth->execute();

    $sth = $dbh->prepare (qq{load data local infile '$binFile' into table
$binTable ignore 1 lines});    $sth->execute();

    $sth = $dbh->prepare (qq{alter table $binTable add column anySNPs enum('y',
'n') default 'n'});  $sth->execute();

    print ("A table with the bins of size $realBin bp has been made at " .
scalar(localtime) . ".\n");
    print LOG ("A table with the bins of size $realBin bp has been made at " .
scalar(localtime) . ".\n");


    ## work on each statistic ##

    foreach $curT (@tables) {
        $inTable    = ("Table" . $curT);
        $valTable   = ("SmallTable" . $curT);

        $sth     =     $dbh->prepare     (qq{drop     table     if     exists
$valTable});   $sth->execute();
```

```
        $sth = $dbh->prepare (qq{create table $valTable select * from $inTable
where negLogVal >= $logValue});    $sth->execute();

        $sth = $dbh->prepare (qq{alter table $valTable add column binStart
int});    $sth->execute();
        $sth = $dbh->prepare (qq{alter table $valTable add column binEnd
int});  $sth->execute();
        $sth = $dbh->prepare (qq{alter table $valTable add index index1(binStart,
binEnd)}); $sth->execute();

        $sth = $dbh->prepare (qq{update $binTable b, $valTable v set v.binStart
= b.binStart where v.chromosome = b.chromosome and v.position between b.binStart
and b.binEnd}); $sth->execute();
        $sth = $dbh->prepare (qq{update $binTable b, $valTable v set v.binEnd
= b.binEnd where v.chromosome = b.chromosome and v.position between b.binStart
and b.binEnd}); $sth->execute();

        $sth = $dbh->prepare (qq{alter table $valTable add column fullLoc
varchar(40)}); $sth->execute();
        $sth = $dbh->prepare (qq{update $valTable set fullLoc =
concat(chromosome, "_", binStart, "_", binEnd)});    $sth->execute();

        print ("\tUpdated the bin positions for $inTable at " .
scalar(localtime) . ".\n");
        print LOG ("\tUpdated the bin positions for $inTable at " .
scalar(localtime) . ".\n");
    }


    ## generate output ##

    my $ST2     = ("SmallTableS2_ihs_resistant");
    my $ST3     = ("SmallTableS3_ihs_resilient");
    my $ST4     = ("SmallTableS4_xpehh");
    my $ST5     = ("SmallTableS5_rsb");

    $sth = $dbh->prepare (qq{drop table if exists $selectedT});  $sth->execute();
    $sth = $dbh->prepare (qq{drop table if exists $selectedT1}); $sth->execute();
    $sth = $dbh->prepare (qq{create table $selectedT1 select SNPname, chromosome,
position, fullLoc from $ST2 union select SNPname, chromosome, position, fullLoc
from $ST3 union select SNPname, chromosome, position, fullLoc from $ST4 union
select SNPname, chromosome, position, fullLoc from $ST5});   $sth->execute();
    $sth = $dbh->prepare (qq{alter table $selectedT1 add index
index1(SNPname)});    $sth->execute();

    $sth = $dbh->prepare (qq{create table $selectedT select * from $selectedT1
order by chromosome, position});  $sth->execute();
    $sth = $dbh->prepare (qq{alter table $selectedT add index index1(SNPname)});
$sth->execute();

    $sth = $dbh->prepare (qq{drop table if exists $selectedT1}); $sth->execute();

    # overall output #

    my @header  = ('SNPname', 'chromosome', 'position', 'fullLocation');
```

```
    foreach $curT (@tables) {
        $valTable   = ("SmallTable" . $curT);
        $value      = ("val" . $curT);
        $negLogVal  = ("negLogVal" . $curT);

        $sth = $dbh->prepare (qq{alter table $selectedT add column $value
float});   $sth->execute();
        $sth = $dbh->prepare (qq{alter table $selectedT add column $negLogVal
float});   $sth->execute();

        push(@header, $value);
        push(@header, $negLogVal);

        $sth = $dbh->prepare (qq{update $selectedT t, $valTable v set t.$value
= v.value where t.SNPname = v.SNPname});  $sth->execute();
        $sth = $dbh->prepare (qq{update $selectedT t, $valTable v set
t.$negLogVal = v.negLogVal where t.SNPname = v.SNPname});  $sth->execute();
    }

    open (OUT, ">$SNPres") or die ("couldn't open $SNPres: $!\n");

    print OUT (join("\t", @header), "\n");

    $statement = ("select * from $selectedT order by chromosome, position");
    &statementPull ($statement, "\t");

    close OUT;

    print ("Created the overall summary output at " . scalar(localtime) .
".\n");
    print LOG ("Created the overall summary output at " . scalar(localtime) .
".\n");

    # per SNP result set #

    foreach $curT (@tables) {
        $outFile    = ($localRes . $compact . "_for_" . $curT . ".txt");
        $value      = ("val" . $curT);
        $negLogVal  = ("negLogVal" . $curT);

        my @header2 = ('SNPname', 'chromosome', 'position', 'fullLocation',
$value, $negLogVal);

        open (OUT, ">$outFile") or die ("couldn't open $outFile: $!\n");

        print OUT (join("\t", @header2), "\n");

        $statement = ("select SNPname, chromosome, position, fullLoc, $value,
$negLogVal from $selectedT where $negLogVal is not null order by chromosome,
position");
        &statementPull ($statement, "\t");

        close OUT;

        print ("Created the table-specific data for $curT at " .
scalar(localtime) . ".\n");
```

```
        print LOG ("Created the table-specific data for $curT at " .
scalar(localtime) . ".\n");
    }

    # overall per bin #

    $sth = $dbh->prepare (qq{drop table if exists localBins});   $sth->execute();
    $sth = $dbh->prepare (qq{create table localBins select chromosome, fullLoc
from $selectedT});    $sth->execute();

    open (OUT, ">$localBins") or die ("couldn't open $localBins: $!\n");

    $statement = ("select * from localBins group by fullLoc");
    &statementPull ($statement, "\t");

    close OUT;

    $sth = $dbh->prepare (qq{drop table if exists localBins});   $sth->execute();

    open (OUT, ">$sortedBin") or die ("couldn't open $sortedBin: $!\n");

    print OUT (join("\t", @header), "\n\n");

    for ($i = 1; $i <= 26; $i++) {

        print OUT ("Overall results for Chromosome $i\n");

        open (IN1, "<$localBins") or die ("couldn't open $localBins: $!\n");

        while (<IN1>) {
            chomp;
            my ($lChrs, $lBin)  = split;

            if ($lChrs eq $i) {
                $statement = ("select * from $selectedT where fullLoc = '$lBin'
and chromosome = '$lChrs' order by position");
                &statementPull ($statement, "\t");

                print OUT ("\n");
            }
        }

        print OUT ("\n");
    }

    close OUT;

    system "rm $binFile";

    print ("Created the overall summary output by chromosome at " .
scalar(localtime) . ".\n");
    print LOG ("Created the overall summary output by chromosome at " .
scalar(localtime) . ".\n");

    return ($binSize, $logValue);
}
```

```
sub dbConnect {
    $count      = 0;
    $rowcount   = 0;

    $datasource = "DBI:mysql:MasseyWork;mysql_local_infile=1";
    $dbh = DBI->connect($datasource, 'pbiggs', 'orange');
    $querystring = '';
}


sub statementPull {
    ($statement, $joiner) = @_;

    $sth = $dbh->prepare (qq{$statement});   $sth->execute();

    $count++;

    while (my @row_items = $sth->fetchrow_array ()) {
        $rowcount++;
        print OUT (join ("$joiner", @row_items), "\n");
        } unless ($rowcount) {
        print OUT ("No data to display\n");
    }

    return ($statement, $joiner);
}
```

5.1 R code for CNVR

```
chromosome <- c(275612895,248993846,224283230,119255633,107901688,117031472,

100079507,90695168,94726778,86447213,62248096,79100223,83079144,
              62722625,80923592,71719816,72286588,68604602,60464314,51176841,
              50073674,50832532,62330649,42034648,45367442,44077779)
rnames <- c("CNVR")
cnames <- c("Chr1","Chr2","Chr3","Chr4","Chr5","Chr6",
           "Chr7","Chr8","Chr9","Chr10","Chr11","Chr12","Chr13",
           "Chr14","Chr15","Chr16","Chr17","Chr18","Chr19","Chr20",
           "Chr21","Chr22","Chr23","Chr24","Chr25","Chr26")
opar<-par(no.readonly=TRUE)
barplot(chromosome, beside=T, horiz=T, xlab="Chromosome Length (Mbp)",
xlim=c(0,300000000), axes=FALSE, cex.names=0.8, las=1, col=0,names.arg=cnames)
axis(1, at=c(0,30000000,60000000,90000000,120000000,150000000,
                    180000000,210000000,240000000,270000000,300000000),
        labels=c(0,30,60,90,120,150,180,210,240,270,300), las=1)
legend(locator(1), title="Event", c("Loss","Gain","Mix"), pch=22,
pt.bg=c("blue","red","yellow"))
attach(loss)
segments(Start,(Chr-1)*1.2+0.7,End,(Chr-1)*1.2+0.7,col="blue",lty=1,lwd=4)
detach(loss)
attach(gain)
segments(Start,(Chr-1)*1.2+0.7,End,(Chr-1)*1.2+0.7,col="red",lty=1,lwd=4)
detach(gain)
attach(mix)
segments(Start,(Chr-1)*1.2+0.7,End,(Chr-1)*1.2+0.7,col="yellow",lty=1,lwd=4)
detach(mix)
par(opar)
```

5.2 R code for CNVR in each tissue

```
library(ggplot2)
library(gridExtra)
tissue <-c("epididymis", "kidney", "liver", "semitendinosus", "testis",
"thymus", "thyroid")
chromosome_1 <-c(rep(275612895, 7))
chromosome_2 <- c(rep(248993846, 7))
chromosome_3 <- c(rep(224283230, 7))
chromosome_4 <- c(rep(119255633, 7))
chromosome_5 <- c(rep(107901688, 7))
chromosome_6 <- c(rep(117031472, 7))
chromosome_7 <- c(rep(100079507, 7))
chromosome_8 <- c(rep(90695168, 7))
chromosome_9 <- c(rep(94726778, 7))
chromosome_10 <- c(rep(86447213, 7))
chromosome_11 <-c(rep(62248096, 7))
chromosome_12 <- c(rep(79100223, 7))
chromosome_13 <- c(rep(83079144, 7))
chromosome_14 <- c(rep(62722625, 7))
chromosome_15 <- c(rep(80923592, 7))
chromosome_16 <- c(rep(71719816, 7))
chromosome_17 <- c(rep(72286588, 7))
chromosome_18 <- c(rep(68604602, 7))
chromosome_19 <- c(rep(60464314, 7))
chromosome_20 <- c(rep(51176841, 7))
chromosome_21 <-c(rep(50073674, 7))
chromosome_22 <- c(rep(50832532, 7))
chromosome_23 <- c(rep(62330649, 7))
chromosome_24 <- c(rep(42034648, 7))
chromosome_25 <- c(rep(45367442, 7))
chromosome_26 <- c(rep(44077779, 7))
chr1_size <- data.frame(tissue, chromosome_1)
chr2_size <- data.frame(tissue, chromosome_2)
chr3_size <- data.frame(tissue, chromosome_3)
chr4_size <- data.frame(tissue, chromosome_4)
chr5_size <- data.frame(tissue, chromosome_5)
chr6_size <- data.frame(tissue, chromosome_6)
chr7_size <- data.frame(tissue, chromosome_7)
chr8_size <- data.frame(tissue, chromosome_8)
chr9_size <- data.frame(tissue, chromosome_9)
chr10_size <- data.frame(tissue, chromosome_10)
chr11_size <- data.frame(tissue, chromosome_11)
chr12_size <- data.frame(tissue, chromosome_12)
chr13_size <- data.frame(tissue, chromosome_13)
chr14_size <- data.frame(tissue, chromosome_14)
chr15_size <- data.frame(tissue, chromosome_15)
chr16_size <- data.frame(tissue, chromosome_16)
chr17_size <- data.frame(tissue, chromosome_17)
chr18_size <- data.frame(tissue, chromosome_18)
chr19_size <- data.frame(tissue, chromosome_19)
chr20_size <- data.frame(tissue, chromosome_20)
chr21_size <- data.frame(tissue, chromosome_21)
chr22_size <- data.frame(tissue, chromosome_22)
chr23_size <- data.frame(tissue, chromosome_23)
chr24_size <- data.frame(tissue, chromosome_24)
```

```
chr25_size <- data.frame(tissue, chromosome_25)
chr26_size <- data.frame(tissue, chromosome_26)
result <-read.table('result.txt', header = TRUE)
chr1 <- read.table('chr1.txt', header = TRUE)
chr2 <- read.table('chr2.txt', header = TRUE)
chr3 <- read.table('chr3.txt', header = TRUE)
chr4 <- read.table('chr4.txt', header = TRUE)
chr5 <- read.table('chr5.txt', header = TRUE)
chr7 <- read.table('chr7.txt', header = TRUE)
chr8 <- read.table('chr8.txt', header = TRUE)
chr9 <- read.table('chr9.txt', header = TRUE)
chr10 <- read.table('chr10.txt', header = TRUE)
chr11 <- read.table('chr11.txt', header = TRUE)
chr12 <- read.table('chr12.txt', header = TRUE)
chr13 <- read.table('chr13.txt', header = TRUE)
chr14 <- read.table('chr14.txt', header = TRUE)
chr15 <- read.table('chr15.txt', header = TRUE)
chr16 <- read.table('chr16.txt', header = TRUE)
chr17 <- read.table('chr17.txt', header = TRUE)
chr18 <- read.table('chr18.txt', header = TRUE)
chr20 <- read.table('chr20.txt', header = TRUE)
chr21 <- read.table('chr21.txt', header = TRUE)
chr22 <- read.table('chr22.txt', header = TRUE)
chr24 <- read.table('chr24.txt', header = TRUE)
number <- factor(chr1$number)
plot_1 <- ggplot()  +
       theme_bw() +
       geom_bar(stat="identity", data = chr1_size, aes(x = tissue, y =
chromosome_1 ), color = 'black', fill = 'white') +
       geom_point(stat="identity", data = chr1, aes(x = tissue, y = mean),
color=number) +
       coord_flip() +
       theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr2$number)
plot_2 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr2_size, aes(x = tissue, y =
chromosome_2 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr2, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr3$number)
plot_3 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr3_size, aes(x = tissue, y =
chromosome_3 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr3, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))
```

```
number <- factor(chr4$number)
plot_4 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr4_size, aes(x = tissue, y =
chromosome_4 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr4, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr5$number)
plot_5 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr5_size, aes(x = tissue, y =
chromosome_5 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr5, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))


plot_6 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr6_size, aes(x = tissue, y =
chromosome_6 ), color = 'black', fill = 'white') +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr7$number)
plot_7 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr7_size, aes(x = tissue, y =
chromosome_7 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr7, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr8$number)
plot_8 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr8_size, aes(x = tissue, y =
chromosome_8 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr8, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr9$number)
plot_9 <- ggplot()  +
```

```
  theme_bw() +
  geom_bar(stat="identity", data = chr9_size, aes(x = tissue, y =
chromosome_9 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr9, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr10$number)
plot_10 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr10_size, aes(x = tissue, y =
chromosome_10 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr10, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

grid.arrange(plot_1, plot_2, plot_3, plot_4, plot_5, plot_6, plot_7, plot_8,
plot_9, plot_10, ncol=2)
# chr11-chr20
number <- factor(chr11$number)
plot_11 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr11_size, aes(x = tissue, y =
chromosome_11 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr11, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr12$number)
plot_12 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr12_size, aes(x = tissue, y =
chromosome_12 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr12, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr13$number)
plot_13 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr13_size, aes(x = tissue, y =
chromosome_13 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr13, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))
```

```
number <- factor(chr14$number)
plot_14 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr14_size, aes(x = tissue, y =
chromosome_14 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr14, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr15$number)
plot_15 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr15_size, aes(x = tissue, y =
chromosome_15 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr15, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr16$number)
plot_16 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr16_size, aes(x = tissue, y =
chromosome_16 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr16, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr17$number)
plot_17 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr17_size, aes(x = tissue, y =
chromosome_17 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr17, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr18$number)
plot_18 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr18_size, aes(x = tissue, y =
chromosome_18 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr18, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))
```

```
plot_19 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr19_size, aes(x = tissue, y =
chromosome_19 ), color = 'black', fill = 'white') +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr20$number)
plot_20 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr20_size, aes(x = tissue, y =
chromosome_20 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr20, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

grid.arrange(plot_11, plot_12, plot_13, plot_14, plot_15, plot_16, plot_17,
plot_18, plot_19, plot_20, ncol=2)

# chr21-chr26
number <- factor(chr21$number)
plot_21 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr21_size, aes(x = tissue, y =
chromosome_21 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr21, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr22$number)
plot_22 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr22_size, aes(x = tissue, y =
chromosome_22 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr22, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))


plot_23 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr23_size, aes(x = tissue, y =
chromosome_23 ), color = 'black', fill = 'white') +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

number <- factor(chr24$number)
plot_24 <- ggplot()  +
```

```
  theme_bw() +
  geom_bar(stat="identity", data = chr24_size, aes(x = tissue, y =
chromosome_24 ), color = 'black', fill = 'white') +
  geom_point(stat="identity", data = chr24, aes(x = tissue, y = mean),
color=number) +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))


plot_25 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr25_size, aes(x = tissue, y =
chromosome_25 ), color = 'black', fill = 'white') +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))


plot_26 <- ggplot()  +
  theme_bw() +
  geom_bar(stat="identity", data = chr26_size, aes(x = tissue, y =
chromosome_26 ), color = 'black', fill = 'white') +
  coord_flip() +
  theme(panel.border=element_blank()) +
scale_y_continuous(limits=c(0,275612895))

grid.arrange(plot_21, plot_22, plot_23, plot_24, plot_25, plot_26, ncol=2)
```

6.1 Code for fastPHASE_v1.4

```
#!/bin/bash
#SBATCH -J haplotype_chr1
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr1 plink.chr-1.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr2
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr2 plink.chr-2.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr3
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr3 plink.chr-3.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr4
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr4 plink.chr-4.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr5
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr5 plink.chr-5.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr6
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
```

```
srun fastPHASE -usubpop.txt -ochr6 plink.chr-6.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr7
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr7 plink.chr-7.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr8
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr8 plink.chr-8.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr9
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr9 plink.chr-9.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr10
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr10 plink.chr-10.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr11
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr11 plink.chr-11.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr12
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr12 plink.chr-12.recode.phase.inp
#!/bin/bash
```

```
#SBATCH -J haplotype_chr13
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr13 plink.chr-13.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr14
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr14 plink.chr-14.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr15
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr15 plink.chr-15.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr16
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr16 plink.chr-16.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr17
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr17 plink.chr-17.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr18
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr18 plink.chr-18.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr19
#SBATCH -A nesi00248
```

```
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr19 plink.chr-19.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr20
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr20 plink.chr-20.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr21
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr21 plink.chr-21.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr22
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr22 plink.chr-22.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr23
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr23 plink.chr-23.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr24
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr24 plink.chr-24.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr25
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
```

```
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr25 plink.chr-25.recode.phase.inp
#!/bin/bash
#SBATCH -J haplotype_chr26
#SBATCH -A nesi00248
#SBATCH --time=96:00:00
#SBATCH --mem-per-cpu=4G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH -C sb
srun fastPHASE -usubpop.txt -ochr26 plink.chr-26.recode.phase.inp
```

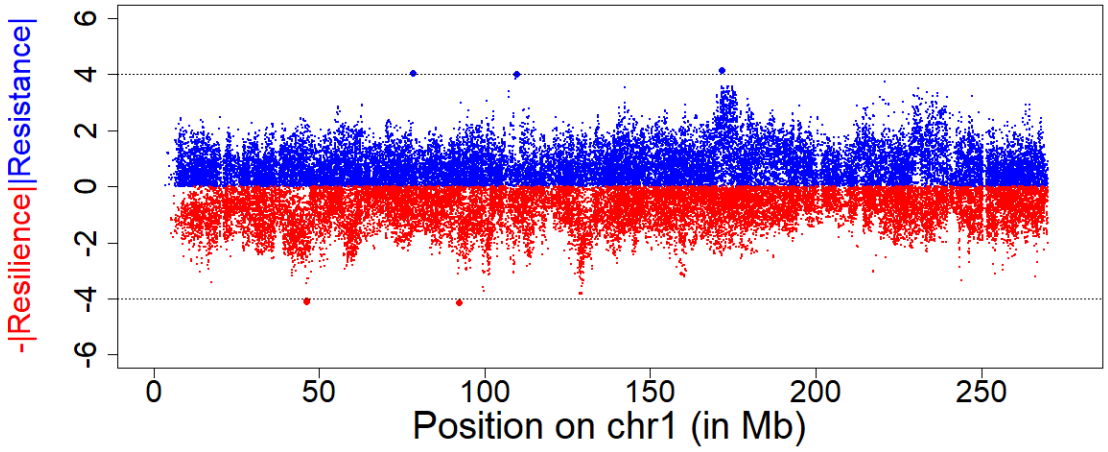
6.2 Code for selection signature regions

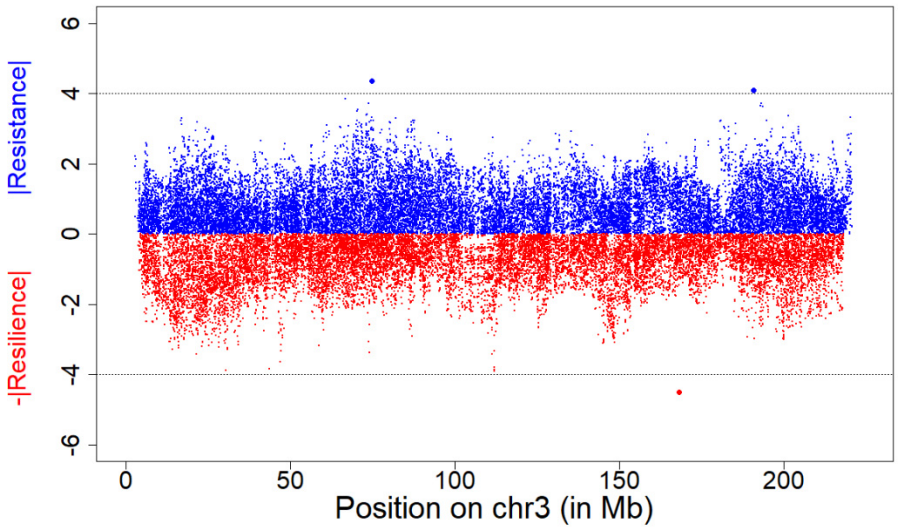
```
output_file = open('result', 'w')
dict_resilience = {}
dict_resistance = {}
dict_rsb = {}
dict_xpehh = {}
start_point = 0
for end_point in range(500000, 83500001, 250000):
    print(end_point)
    list = []
    input_file = open('ihs_resilient_all.txt')
    for i in range(380000):
        line = input_file.readline().strip().split()
        if (line[1] == '13') and (int(line[2]) in range(start_point,
end_point)):
            list.append((int(line[2]), float(line[3]), float(line[4])))
    dict_resilience[end_point] = list
    start_point = start_point + 250000
print('dict_resilience_set up')
start_point = 0
for end_point in range(500000, 83500001, 250000):
    print(end_point)
    list = []
    input_file = open('ihs_resistant_all.txt')
    for i in range(380000):
        line = input_file.readline().strip().split()
        if (line[1] == '13') and (int(line[2]) in range(start_point,
end_point)):
            list.append((int(line[2]), float(line[3]), float(line[4])))
    dict_resistance[end_point] = list
    start_point = start_point + 250000
print('dict_resistance_set up')
start_point = 0
for end_point in range(500000, 83500001, 250000):
    print(end_point)
    list = []
    input_file = open('rsb_all.txt')
    for i in range(380000):
        line = input_file.readline().strip().split()
        if (line[1] == '13') and (int(line[2]) in range(start_point,
end_point)):
            list.append((int(line[2]), float(line[3]), float(line[4])))
    dict_rsb[end_point] = list
    start_point = start_point + 250000
print('dict_rsb_set up')
start_point = 0
for end_point in range(500000, 83500001, 250000):
    print(end_point)
    list = []
    input_file = open('xpehh_all.txt')
    for i in range(380000):
        line = input_file.readline().strip().split()
        if (line[1] == '13') and (int(line[2]) in range(start_point,
end_point)):
            list.append((int(line[2]), float(line[3]), float(line[4])))
```

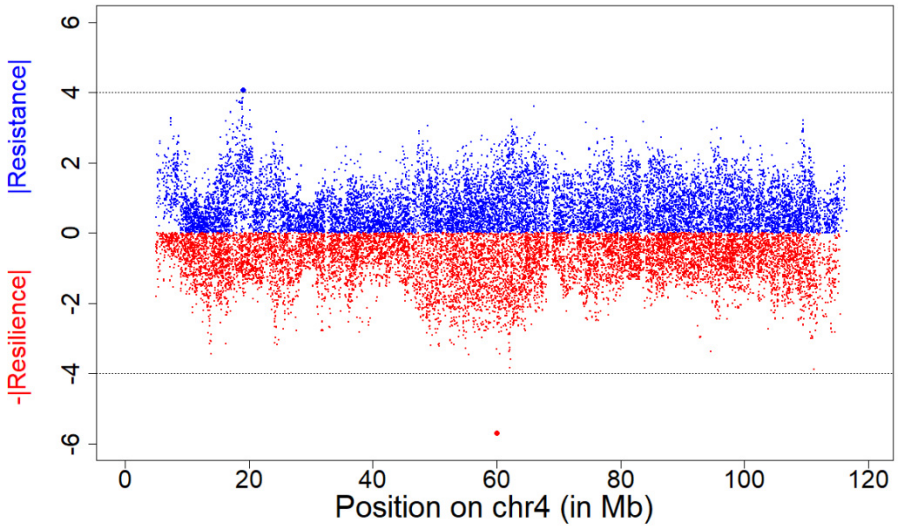
```
    dict_xpehh[end_point] = list
    start_point = start_point + 250000
print('dict_xpehh_set up')
start_point = 0
for end_point in range(500000, 83500001, 250000):
    print(end_point)
    position = (start_point, end_point)
    start_point = start_point + 250000
    list_resistance = []
    for i in sorted(dict_resistance[end_point], key=lambda x: x[1]):
        if i != []:
            list_resistance.append(i)
        peak_resistance = list_resistance[0]
    list_resilience = []
    for i in sorted(dict_resilience[end_point], key=lambda x: x[1]):
        if i != []:
            list_resilience.append(i)
        peak_resilience = list_resilience[0]
    list_rsb = []
    for i in sorted(dict_rsb[end_point], key=lambda x: x[1]):
        if i != []:
            list_rsb.append(i)
        peak_rsb = list_rsb[-1]
    list_xpehh = []
    for i in sorted(dict_xpehh[end_point], key=lambda x: x[1]):
        if i != []:
            list_xpehh.append(i)
        peak_xpehh = list_xpehh[-1]
    peak_resistance_position = peak_resistance[0]
    peak_resilience_position = peak_resilience[0]
    peak_rsb_position = peak_rsb[0]
    peak_xpehh_position = peak_xpehh[0]
    peak_resistance_value = peak_resistance[1]
    peak_resilience_value = peak_resilience[1]
    peak_rsb_value = peak_rsb[1]
    peak_xpehh_value = peak_xpehh[1]
    resistance_p_4 = 0
    for k in sorted(dict_resistance[end_point], key=lambda x:x[1]):
        if k != []:
            if k[2] >= 4:
                resistance_p_4 = resistance_p_4 + 1
    resilience_p_4 = 0
    for k in sorted(dict_resilience[end_point], key=lambda x:x[1]):
        if k != []:
            if k[2] >= 4:
                resilience_p_4 = resilience_p_4 + 1
    rsb_p_4 = 0
    for k in sorted(dict_rsb[end_point], key=lambda x: x[1]):
        if k != []:
            if k[2] >= 4:
                rsb_p_4 = rsb_p_4 + 1
    xpehh_p_4 = 0
    for k in sorted(dict_xpehh[end_point], key=lambda x:x[1]):
        if k != []:
            if k[2] >= 4:
                xpehh_p_4 = xpehh_p_4 + 1
```

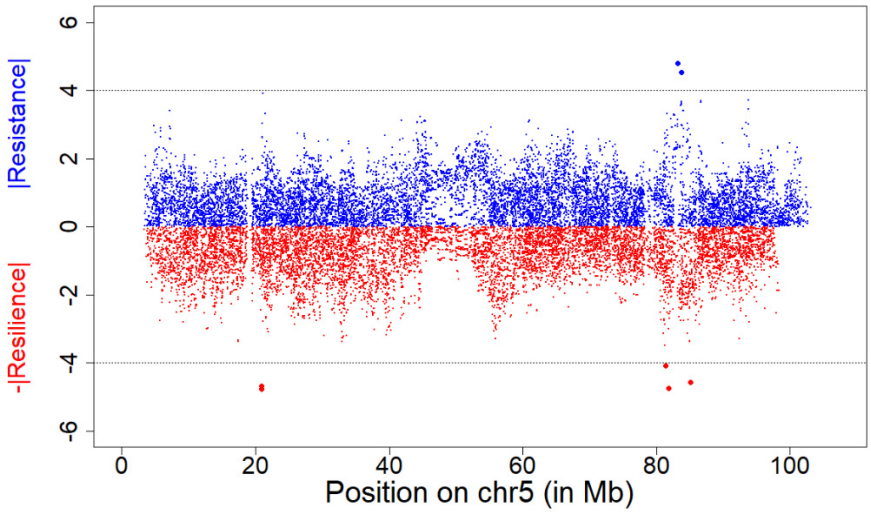
```
    if resilience_p_4 >= 2 or resistance_p_4 >= 2 or rsb_p_4 >= 2 or
xpehh_p_4 >= 2:
        output_file.write(str(start_point) + ' ' + str(end_point) + ' ' +
'iHS_resistance' + str(peak_resistance_position) \
                  + ' ' + str(peak_resistance_value) + ' ' +
str(resistance_p_4) + '\n')
        output_file.write(str(start_point) + ' ' + str(end_point) + ' ' +
'iHS_resilience' + str(peak_resilience_position) \
                  + ' ' + str(peak_resilience_value) + ' ' +
str(resilience_p_4) + '\n')
        output_file.write(str(start_point) + ' ' + str(end_point) + ' ' +
'rsb' + str(peak_rsb_position) \
                  + ' ' + str(peak_rsb_value) + ' ' + str(rsb_p_4) + '\n')
        output_file.write(str(start_point) + ' ' + str(end_point) + ' ' +
'xpehh' + str(peak_xpehh_position) \
                  + ' ' + str(peak_xpehh_value) + ' ' + str(xpehh_p_4) +
'\n')
output_file.close()
```

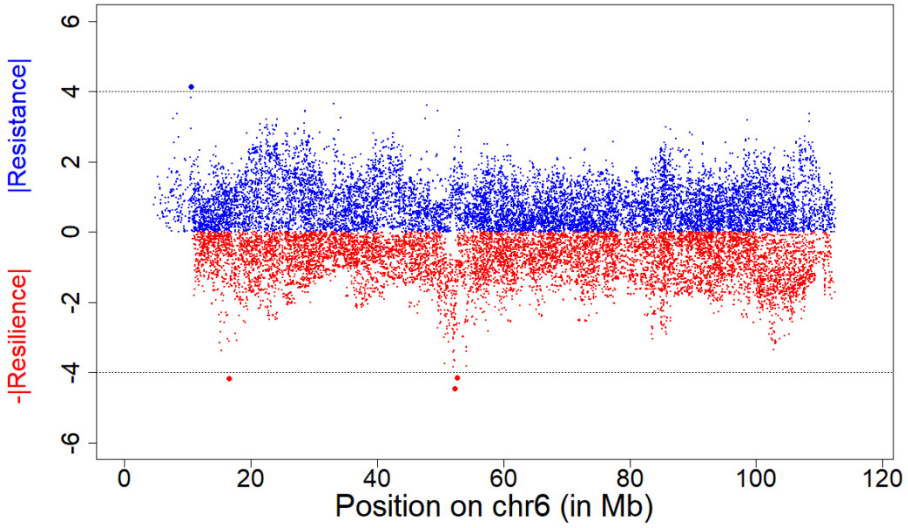
6.3 iHS plots in the two lines

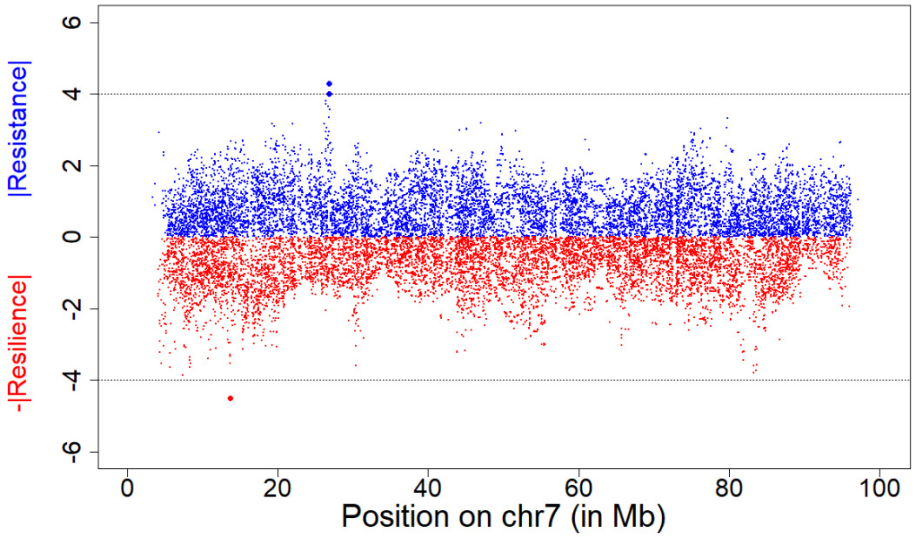


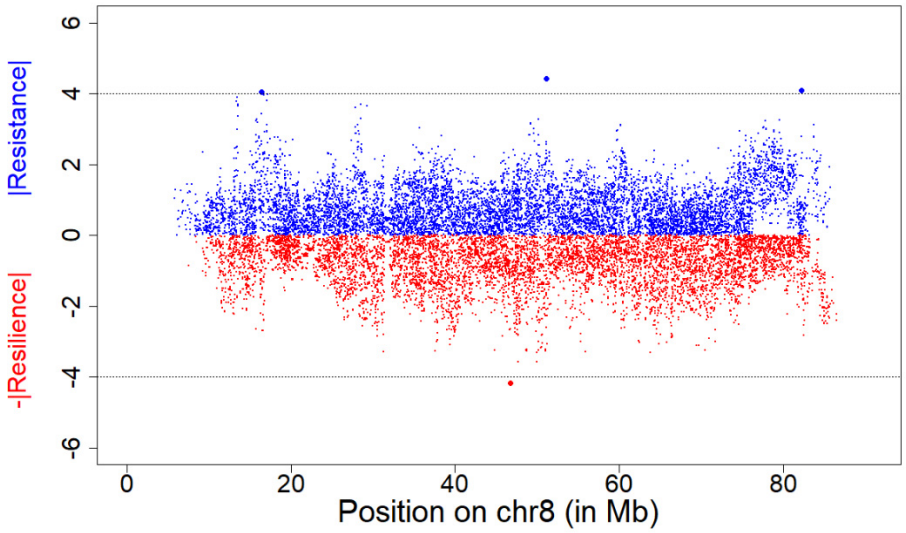


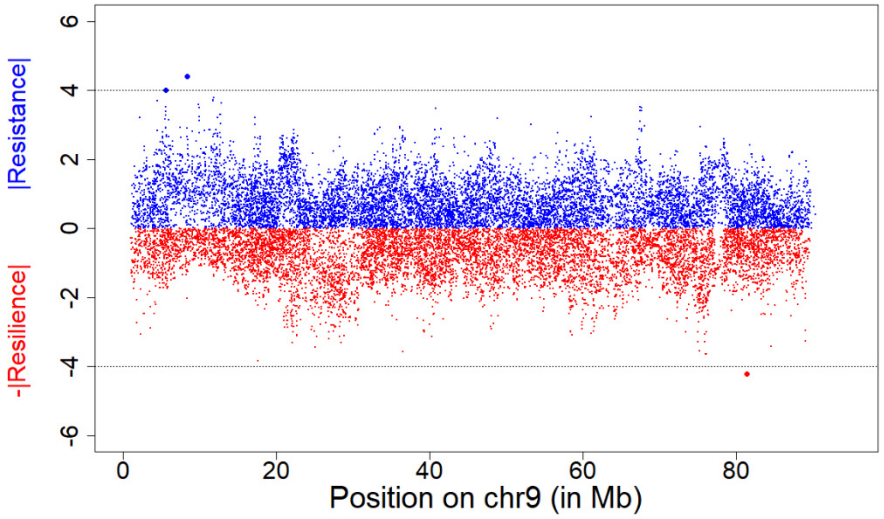


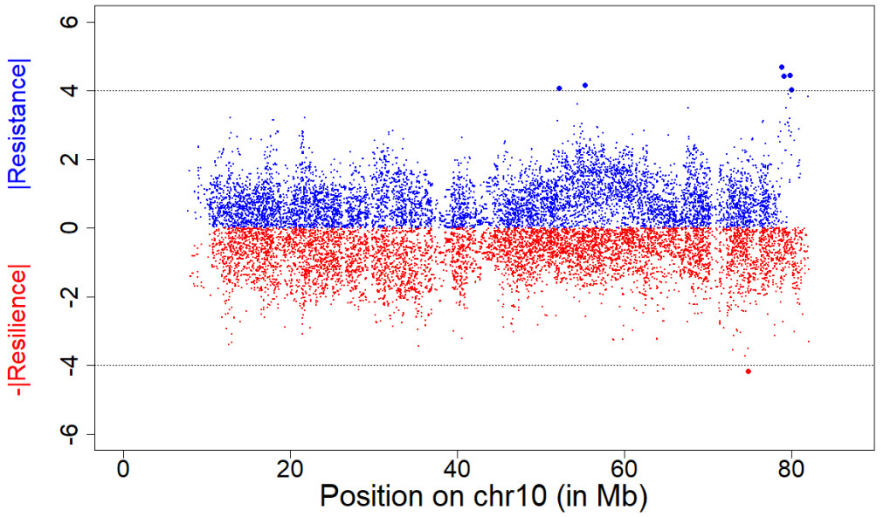

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
100
Position on chr5 (in Mb)

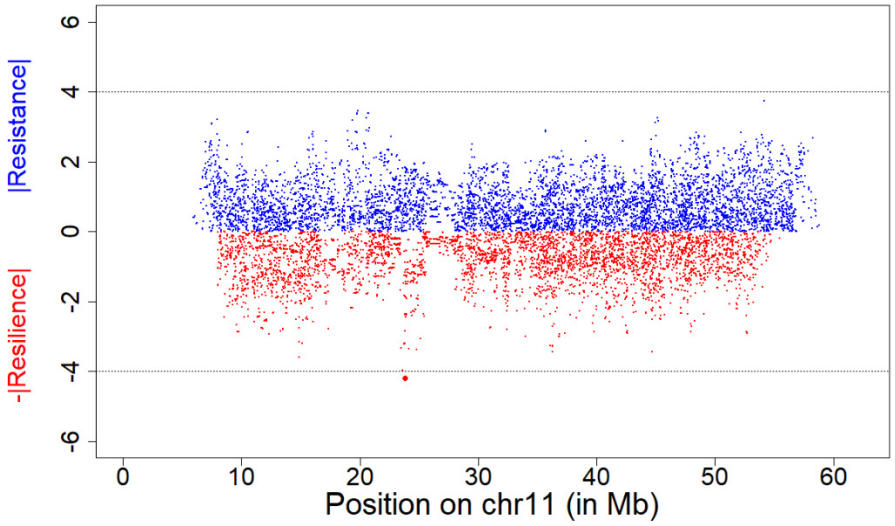

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
100
120
Position on chr6 (in Mb)

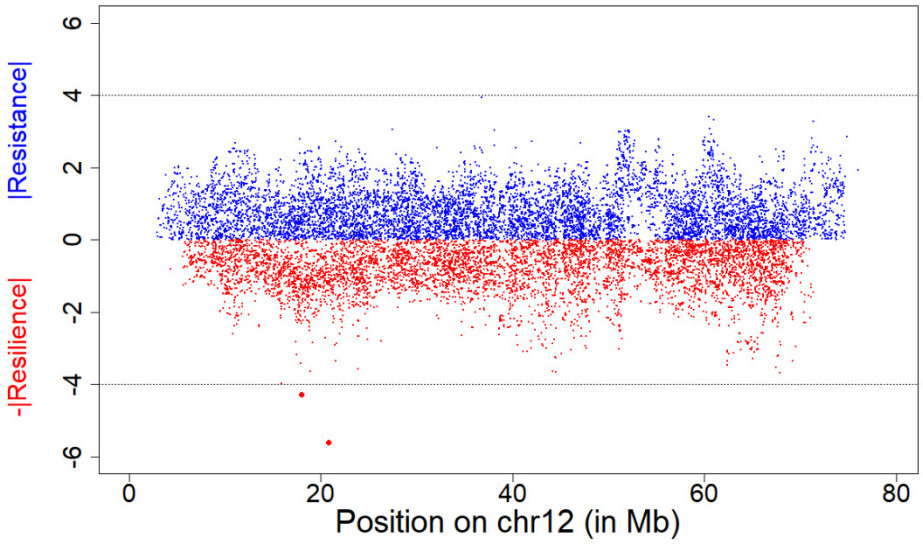

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
100
Position on chr7 (in Mb)

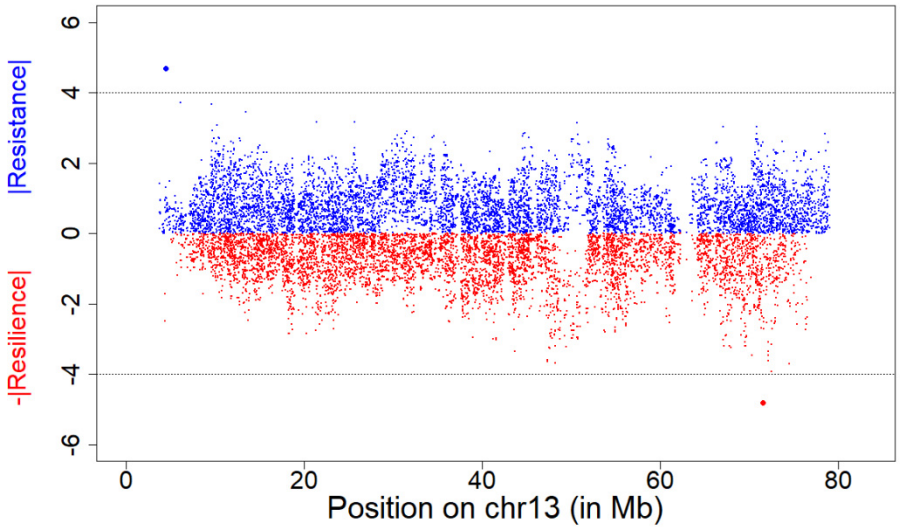

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
Position on chr8 (in Mb)

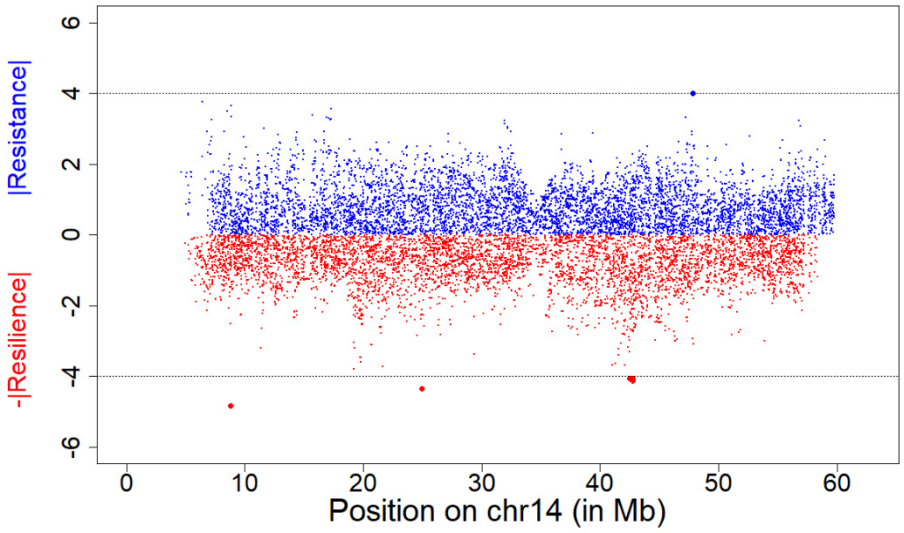

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
Position on chr9 (in Mb)

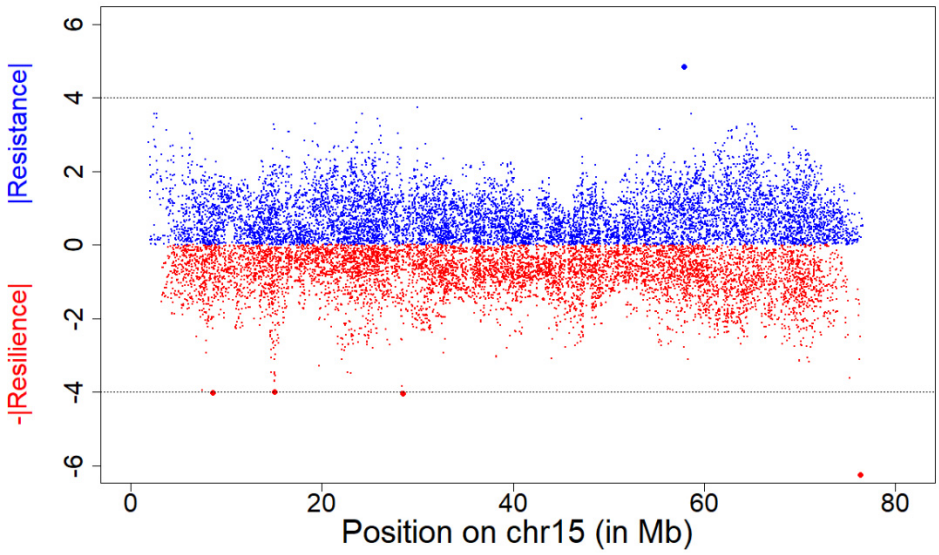

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
Position on chr10 (in Mb)

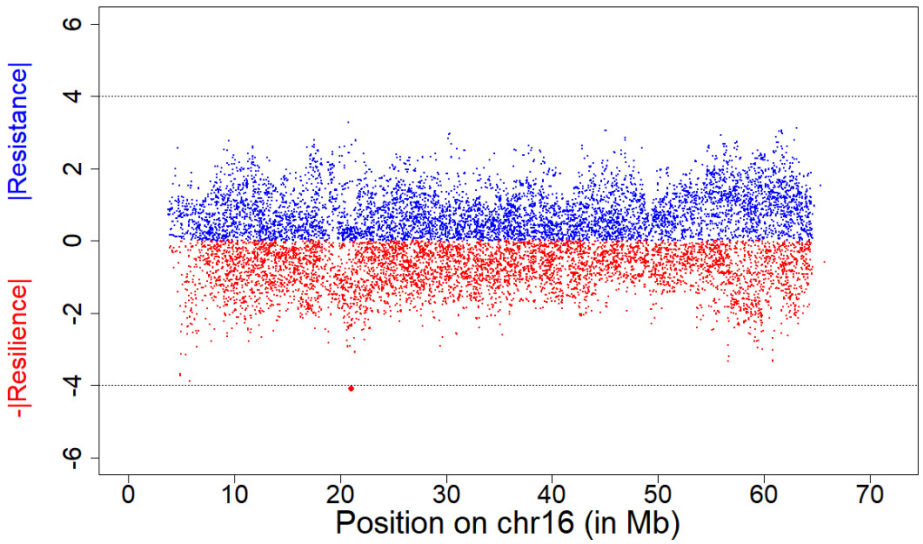

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
60
Position on chr11 (in Mb)

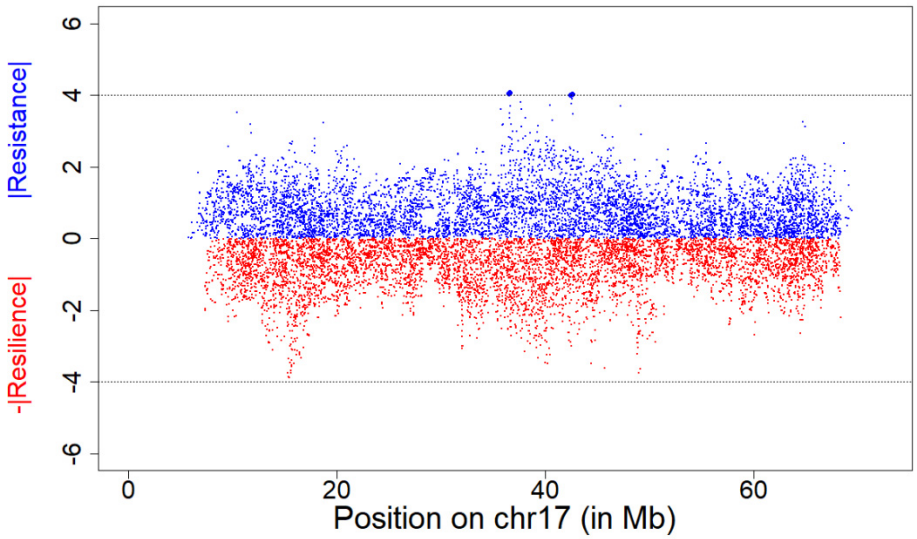

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
Position on chr12 (in Mb)

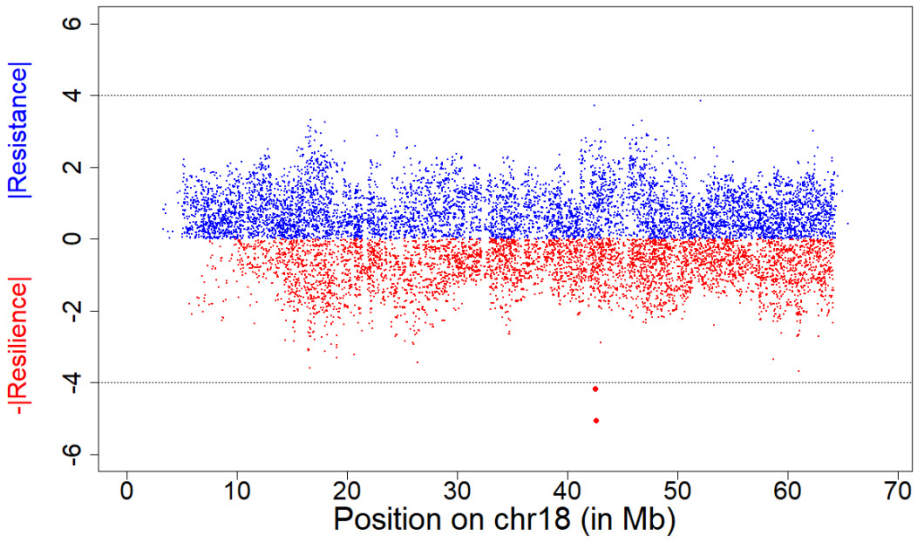

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
Position on chr13 (in Mb)

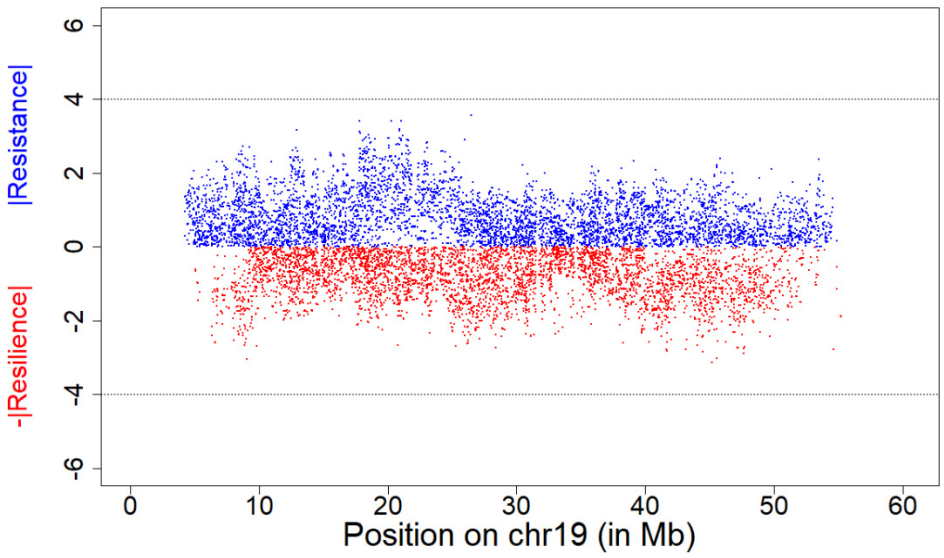

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
60
Position on chr14 (in Mb)

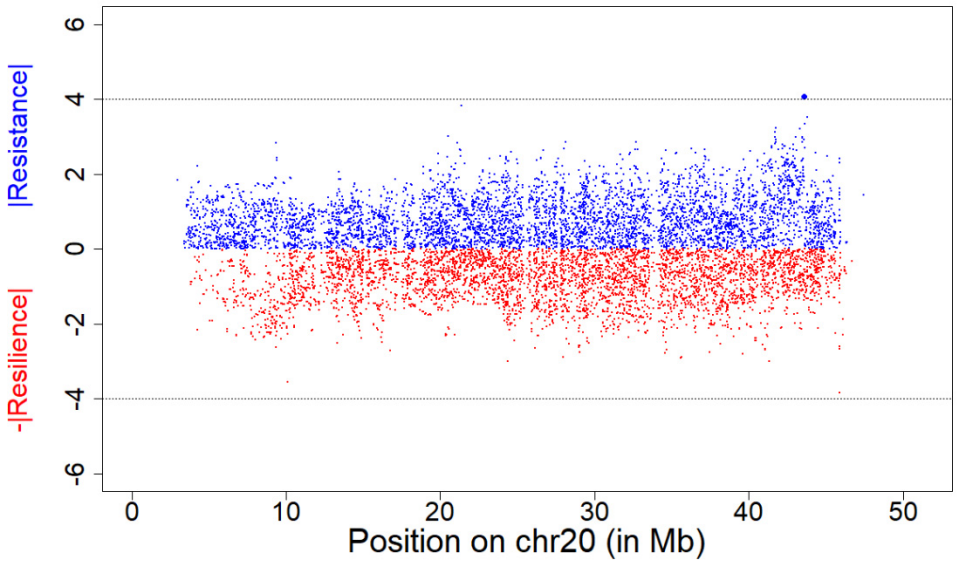

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
80
Position on chr15 (in Mb)

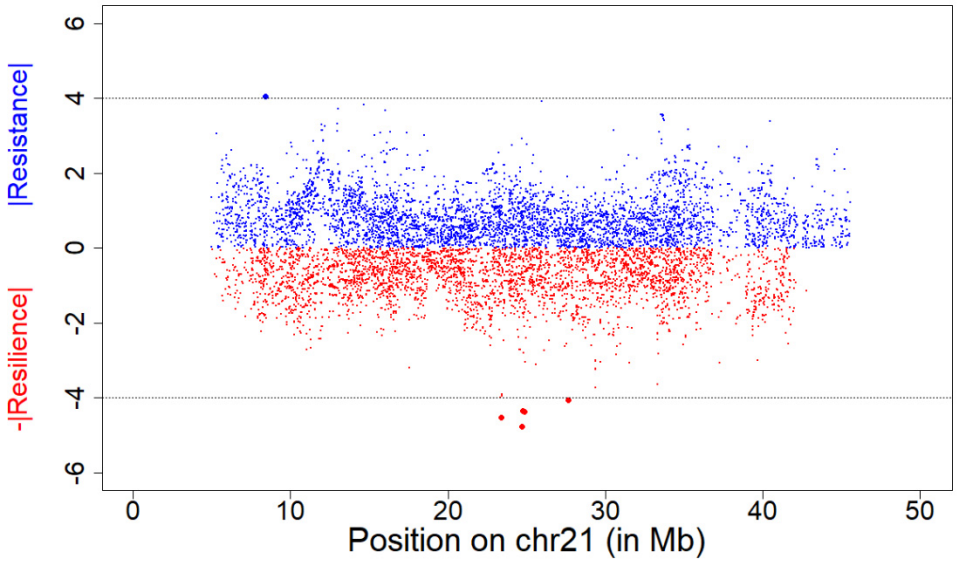

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
60
70
Position on chr16 (in Mb)

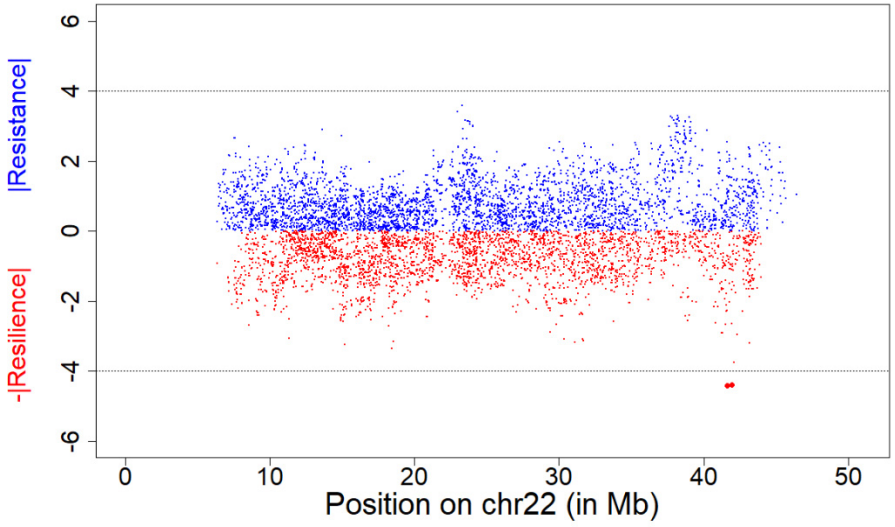

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
20
40
60
Position on chr17 (in Mb)

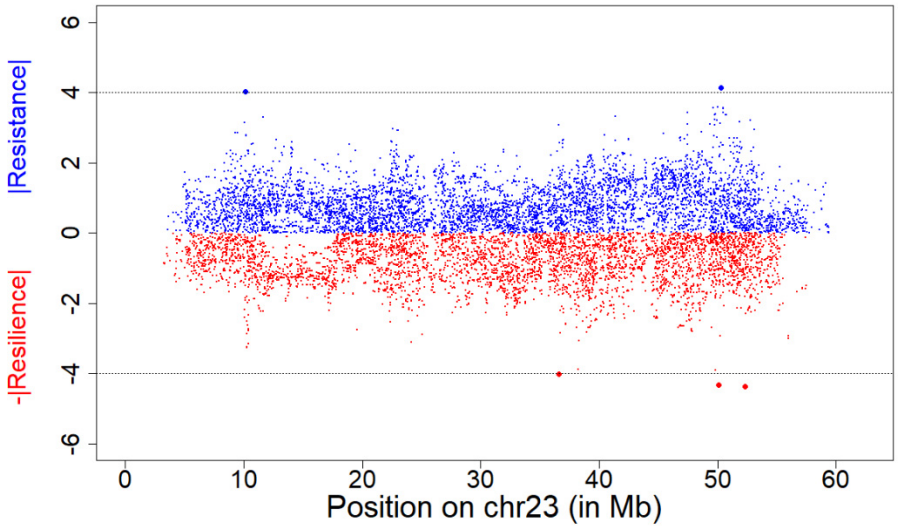

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
60
70
Position on chr18 (in Mb)

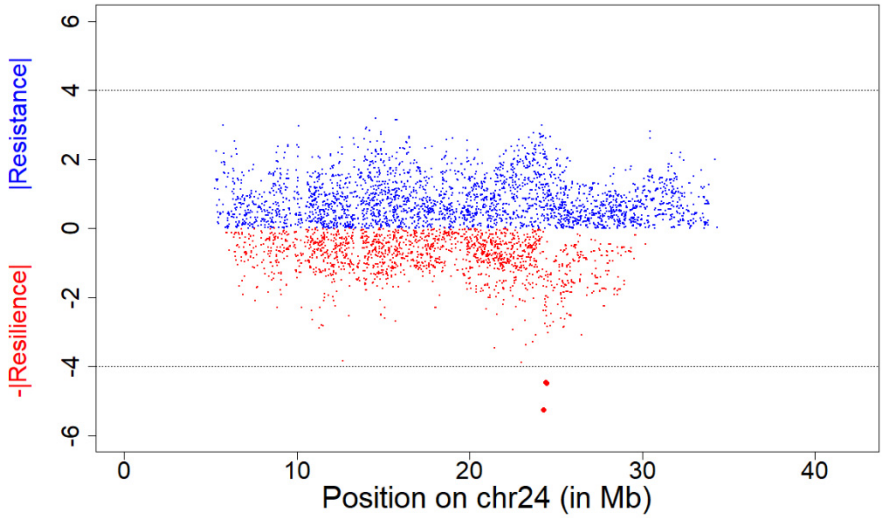

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
60
Position on chr19 (in Mb)

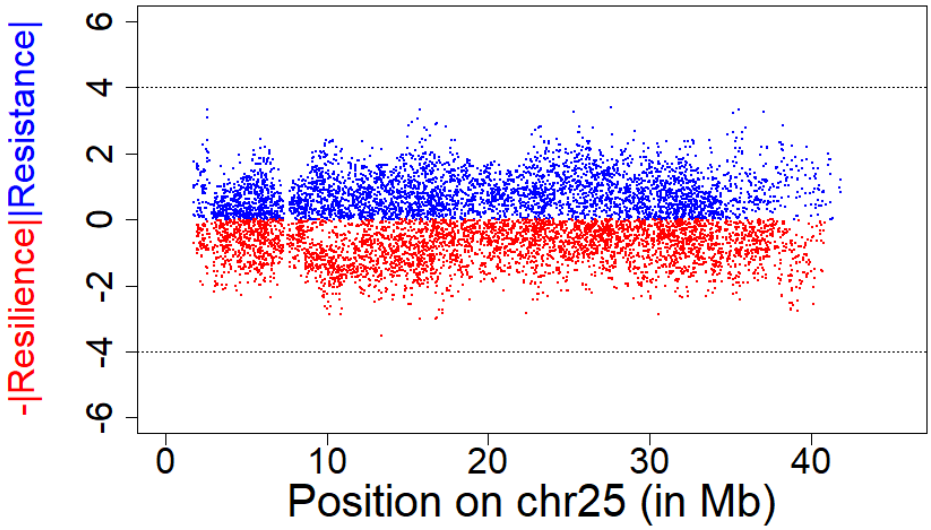

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
Position on chr20 (in Mb)

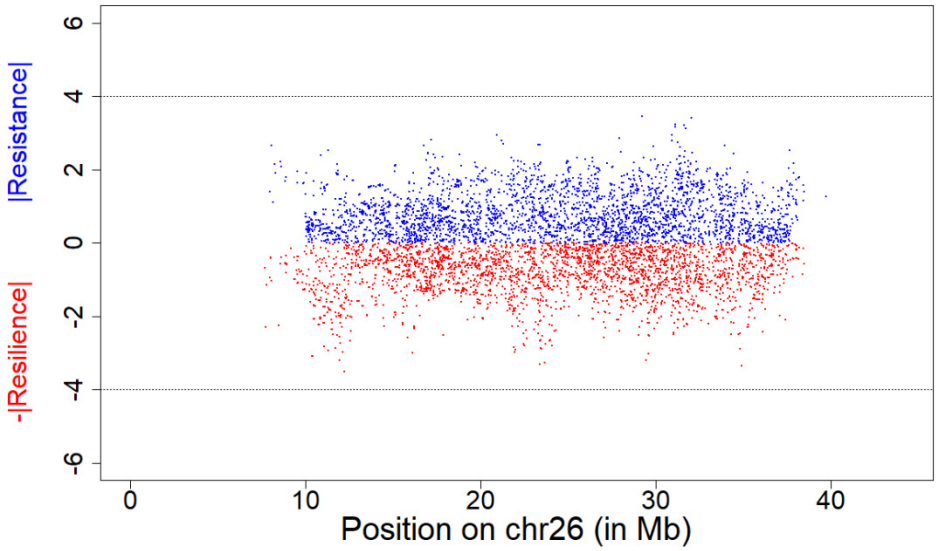

|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
Position on chr21 (in Mb)


|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
Position on chr22 (in Mb)


|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
50
60
Position on chr23 (in Mb)


|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
Position on chr24 (in Mb)


-|Resilience||Resistance|
6
4
2
0
-2
-4
-6
0
10
20
30
40
Position on chr25 (in Mb)


|Resistance|
-|Resilience|
6
4
2
0
-2
-4
-6
0
10
20
30
40
Position on chr26 (in Mb)

6.4 R code for EHH and EHHS

```
#input data from fastphase results
hap_chr1_resilient<-
data2haplohh(hap_file="chr1_hapguess_switch.out",map_file="chr1_map.inp",
                              popsel=1,chr.name=1,recode.allele=TRUE)
hap_chr1_resistant<-
data2haplohh(hap_file="chr1_hapguess_switch.out",map_file="chr1_map.inp",
                              popsel=2,chr.name=1,recode.allele=TRUE)

hap_chr2_resilient<-
data2haplohh(hap_file="chr2_hapguess_switch.out",map_file="chr2_map.inp",
                             popsel=1,chr.name=2,recode.allele=TRUE)
hap_chr2_resistant<-
data2haplohh(hap_file="chr2_hapguess_switch.out",map_file="chr2_map.inp",
                             popsel=2,chr.name=2,recode.allele=TRUE)

hap_chr3_resilient<-
data2haplohh(hap_file="chr3_hapguess_switch.out",map_file="chr3_map.inp",
                              popsel=1,chr.name=3,recode.allele=TRUE)
hap_chr3_resistant<-
data2haplohh(hap_file="chr3_hapguess_switch.out",map_file="chr3_map.inp",
                              popsel=2,chr.name=3,recode.allele=TRUE)

hap_chr4_resilient<-
data2haplohh(hap_file="chr4_hapguess_switch.out",map_file="chr4_map.inp",
                              popsel=1,chr.name=4,recode.allele=TRUE)
hap_chr4_resistant<-
data2haplohh(hap_file="chr4_hapguess_switch.out",map_file="chr4_map.inp",
                              popsel=2,chr.name=4,recode.allele=TRUE)

hap_chr5_resilient<-
data2haplohh(hap_file="chr5_hapguess_switch.out",map_file="chr5_map.inp",
                              popsel=1,chr.name=5,recode.allele=TRUE)
hap_chr5_resistant<-
data2haplohh(hap_file="chr5_hapguess_switch.out",map_file="chr5_map.inp",
                              popsel=2,chr.name=5,recode.allele=TRUE)

hap_chr6_resilient<-
data2haplohh(hap_file="chr6_hapguess_switch.out",map_file="chr6_map.inp",
                              popsel=1,chr.name=6,recode.allele=TRUE)
hap_chr6_resistant<-
data2haplohh(hap_file="chr6_hapguess_switch.out",map_file="chr6_map.inp",
                              popsel=2,chr.name=6,recode.allele=TRUE)

hap_chr7_resilient<-
data2haplohh(hap_file="chr7_hapguess_switch.out",map_file="chr7_map.inp",
                              popsel=1,chr.name=7,recode.allele=TRUE)
hap_chr7_resistant<-
data2haplohh(hap_file="chr7_hapguess_switch.out",map_file="chr7_map.inp",
                              popsel=2,chr.name=7,recode.allele=TRUE)

hap_chr8_resilient<-
data2haplohh(hap_file="chr8_hapguess_switch.out",map_file="chr8_map.inp",
                              popsel=1,chr.name=8,recode.allele=TRUE)
```

```
hap_chr8_resistant<-
data2haplohh(hap_file="chr8_hapguess_switch.out",map_file="chr8_map.inp",
                              popsel=2,chr.name=8,recode.allele=TRUE)

hap_chr9_resilient<-
data2haplohh(hap_file="chr9_hapguess_switch.out",map_file="chr9_map.inp",
                              popsel=1,chr.name=9,recode.allele=TRUE)
hap_chr9_resistant<-
data2haplohh(hap_file="chr9_hapguess_switch.out",map_file="chr9_map.inp",
                              popsel=2,chr.name=9,recode.allele=TRUE)

hap_chr10_resilient<-
data2haplohh(hap_file="chr10_hapguess_switch.out",map_file="chr10_map.inp",
                              popsel=1,chr.name=10,recode.allele=TRUE)
hap_chr10_resistant<-
data2haplohh(hap_file="chr10_hapguess_switch.out",map_file="chr10_map.inp",
                              popsel=2,chr.name=10,recode.allele=TRUE)

hap_chr11_resilient<-
data2haplohh(hap_file="chr11_hapguess_switch.out",map_file="chr11_map.inp",
                              popsel=1,chr.name=11,recode.allele=TRUE)
hap_chr11_resistant<-
data2haplohh(hap_file="chr11_hapguess_switch.out",map_file="chr11_map.inp",
                              popsel=2,chr.name=11,recode.allele=TRUE)

hap_chr12_resilient<-
data2haplohh(hap_file="chr12_hapguess_switch.out",map_file="chr12_map.inp",
                              popsel=1,chr.name=12,recode.allele=TRUE)
hap_chr12_resistant<-
data2haplohh(hap_file="chr12_hapguess_switch.out",map_file="chr12_map.inp",
                              popsel=2,chr.name=12,recode.allele=TRUE)

hap_chr13_resilient<-
data2haplohh(hap_file="chr13_hapguess_switch.out",map_file="chr13_map.inp",
                              popsel=1,chr.name=13,recode.allele=TRUE)
hap_chr13_resistant<-
data2haplohh(hap_file="chr13_hapguess_switch.out",map_file="chr13_map.inp",
                              popsel=2,chr.name=13,recode.allele=TRUE)

hap_chr14_resilient<-
data2haplohh(hap_file="chr14_hapguess_switch.out",map_file="chr14_map.inp",
                              popsel=1,chr.name=14,recode.allele=TRUE)
hap_chr14_resistant<-
data2haplohh(hap_file="chr14_hapguess_switch.out",map_file="chr14_map.inp",
                              popsel=2,chr.name=14,recode.allele=TRUE)

hap_chr15_resilient<-
data2haplohh(hap_file="chr15_hapguess_switch.out",map_file="chr15_map.inp",
                              popsel=1,chr.name=15,recode.allele=TRUE)
hap_chr15_resistant<-
data2haplohh(hap_file="chr15_hapguess_switch.out",map_file="chr15_map.inp",
                              popsel=2,chr.name=15,recode.allele=TRUE)

hap_chr16_resilient<-
data2haplohh(hap_file="chr16_hapguess_switch.out",map_file="chr16_map.inp",
                              popsel=1,chr.name=16,recode.allele=TRUE)
```

```
hap_chr16_resistant<-
data2haplohh(hap_file="chr16_hapguess_switch.out",map_file="chr16_map.inp",
                              popsel=2,chr.name=16,recode.allele=TRUE)

hap_chr17_resilient<-
data2haplohh(hap_file="chr17_hapguess_switch.out",map_file="chr17_map.inp",
                              popsel=1,chr.name=17,recode.allele=TRUE)
hap_chr17_resistant<-
data2haplohh(hap_file="chr17_hapguess_switch.out",map_file="chr17_map.inp",
                              popsel=2,chr.name=17,recode.allele=TRUE)

hap_chr18_resilient<-
data2haplohh(hap_file="chr18_hapguess_switch.out",map_file="chr18_map.inp",
                              popsel=1,chr.name=18,recode.allele=TRUE)
hap_chr18_resistant<-
data2haplohh(hap_file="chr18_hapguess_switch.out",map_file="chr18_map.inp",
                              popsel=2,chr.name=18,recode.allele=TRUE)

hap_chr19_resilient<-
data2haplohh(hap_file="chr19_hapguess_switch.out",map_file="chr19_map.inp",
                              popsel=1,chr.name=19,recode.allele=TRUE)
hap_chr19_resistant<-
data2haplohh(hap_file="chr19_hapguess_switch.out",map_file="chr19_map.inp",
                              popsel=2,chr.name=19,recode.allele=TRUE)

hap_chr20_resilient<-
data2haplohh(hap_file="chr20_hapguess_switch.out",map_file="chr20_map.inp",
                              popsel=1,chr.name=20,recode.allele=TRUE)
hap_chr20_resistant<-
data2haplohh(hap_file="chr20_hapguess_switch.out",map_file="chr20_map.inp",
                              popsel=2,chr.name=20,recode.allele=TRUE)

hap_chr21_resilient<-
data2haplohh(hap_file="chr21_hapguess_switch.out",map_file="chr21_map.inp",
                              popsel=1,chr.name=21,recode.allele=TRUE)
hap_chr21_resistant<-
data2haplohh(hap_file="chr21_hapguess_switch.out",map_file="chr21_map.inp",
                              popsel=2,chr.name=21,recode.allele=TRUE)

hap_chr22_resilient<-
data2haplohh(hap_file="chr22_hapguess_switch.out",map_file="chr22_map.inp",
                              popsel=1,chr.name=22,recode.allele=TRUE)
hap_chr22_resistant<-
data2haplohh(hap_file="chr22_hapguess_switch.out",map_file="chr22_map.inp",
                              popsel=2,chr.name=22,recode.allele=TRUE)

hap_chr23_resilient<-
data2haplohh(hap_file="chr23_hapguess_switch.out",map_file="chr23_map.inp",
                              popsel=1,chr.name=23,recode.allele=TRUE)
hap_chr23_resistant<-
data2haplohh(hap_file="chr23_hapguess_switch.out",map_file="chr23_map.inp",
                              popsel=2,chr.name=23,recode.allele=TRUE)

hap_chr24_resilient<-
data2haplohh(hap_file="chr24_hapguess_switch.out",map_file="chr24_map.inp",
                              popsel=1,chr.name=24,recode.allele=TRUE)
```

```
hap_chr24_resistant<-
data2haplohh(hap_file="chr24_hapguess_switch.out",map_file="chr24_map.inp",
                              popsel=2,chr.name=24,recode.allele=TRUE)

hap_chr25_resilient<-
data2haplohh(hap_file="chr25_hapguess_switch.out",map_file="chr25_map.inp",
                              popsel=1,chr.name=25,recode.allele=TRUE)
hap_chr25_resistant<-
data2haplohh(hap_file="chr25_hapguess_switch.out",map_file="chr25_map.inp",
                              popsel=2,chr.name=25,recode.allele=TRUE)

hap_chr26_resilient<-
data2haplohh(hap_file="chr26_hapguess_switch.out",map_file="chr26_map.inp",
                              popsel=1,chr.name=26,recode.allele=TRUE)
hap_chr26_resistant<-
data2haplohh(hap_file="chr26_hapguess_switch.out",map_file="chr26_map.inp",
                              popsel=2,chr.name=26,recode.allele=TRUE)
#scan data to get haplotype data,such as iHH and iES
hap_chr1_resilient.scan<-scan_hh(hap_chr1_resilient)
hap_chr1_resistant.scan<-scan_hh(hap_chr1_resistant)

hap_chr2_resilient.scan<-scan_hh(hap_chr2_resilient)
hap_chr2_resistant.scan<-scan_hh(hap_chr2_resistant)

hap_chr3_resilient.scan<-scan_hh(hap_chr3_resilient)
hap_chr3_resistant.scan<-scan_hh(hap_chr3_resistant)

hap_chr4_resilient.scan<-scan_hh(hap_chr4_resilient)
hap_chr4_resistant.scan<-scan_hh(hap_chr4_resistant)

hap_chr5_resilient.scan<-scan_hh(hap_chr5_resilient)
hap_chr5_resistant.scan<-scan_hh(hap_chr5_resistant)

hap_chr6_resilient.scan<-scan_hh(hap_chr6_resilient)
hap_chr6_resistant.scan<-scan_hh(hap_chr6_resistant)

hap_chr7_resilient.scan<-scan_hh(hap_chr7_resilient)
hap_chr7_resistant.scan<-scan_hh(hap_chr7_resistant)

hap_chr8_resilient.scan<-scan_hh(hap_chr8_resilient)
hap_chr8_resistant.scan<-scan_hh(hap_chr8_resistant)

hap_chr9_resilient.scan<-scan_hh(hap_chr9_resilient)
hap_chr9_resistant.scan<-scan_hh(hap_chr9_resistant)

hap_chr10_resilient.scan<-scan_hh(hap_chr10_resilient)
hap_chr10_resistant.scan<-scan_hh(hap_chr10_resistant)

hap_chr11_resilient.scan<-scan_hh(hap_chr11_resilient)
hap_chr11_resistant.scan<-scan_hh(hap_chr11_resistant)

hap_chr12_resilient.scan<-scan_hh(hap_chr12_resilient)
hap_chr12_resistant.scan<-scan_hh(hap_chr12_resistant)

hap_chr13_resilient.scan<-scan_hh(hap_chr13_resilient)
hap_chr13_resistant.scan<-scan_hh(hap_chr13_resistant)
```

```
hap_chr14_resilient.scan<-scan_hh(hap_chr14_resilient)
hap_chr14_resistant.scan<-scan_hh(hap_chr14_resistant)

hap_chr15_resilient.scan<-scan_hh(hap_chr15_resilient)
hap_chr15_resistant.scan<-scan_hh(hap_chr15_resistant)

hap_chr16_resilient.scan<-scan_hh(hap_chr16_resilient)
hap_chr16_resistant.scan<-scan_hh(hap_chr16_resistant)

hap_chr17_resilient.scan<-scan_hh(hap_chr17_resilient)
hap_chr17_resistant.scan<-scan_hh(hap_chr17_resistant)

hap_chr18_resilient.scan<-scan_hh(hap_chr18_resilient)
hap_chr18_resistant.scan<-scan_hh(hap_chr18_resistant)

hap_chr19_resilient.scan<-scan_hh(hap_chr19_resilient)
hap_chr19_resistant.scan<-scan_hh(hap_chr19_resistant)

hap_chr20_resilient.scan<-scan_hh(hap_chr20_resilient)
hap_chr20_resistant.scan<-scan_hh(hap_chr20_resistant)

hap_chr21_resilient.scan<-scan_hh(hap_chr21_resilient)
hap_chr21_resistant.scan<-scan_hh(hap_chr21_resistant)

hap_chr22_resilient.scan<-scan_hh(hap_chr22_resilient)
hap_chr22_resistant.scan<-scan_hh(hap_chr22_resistant)

hap_chr23_resilient.scan<-scan_hh(hap_chr23_resilient)
hap_chr23_resistant.scan<-scan_hh(hap_chr23_resistant)

hap_chr24_resilient.scan<-scan_hh(hap_chr24_resilient)
hap_chr24_resistant.scan<-scan_hh(hap_chr24_resistant)

hap_chr25_resilient.scan<-scan_hh(hap_chr25_resilient)
hap_chr25_resistant.scan<-scan_hh(hap_chr25_resistant)

hap_chr26_resilient.scan<-scan_hh(hap_chr26_resilient)
hap_chr26_resistant.scan<-scan_hh(hap_chr26_resistant)


#calculate ihs from scan
hap_chr1_resilient.ihs<-ihh2ihs(hap_chr1_resilient.scan)
write.csv(hap_chr1_resilient.ihs$iHS,"ihs_resilient_chr1.csv")
write.csv(hap_chr1_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r1.csv")
hap_chr1_resistant.ihs<-ihh2ihs(hap_chr1_resistant.scan)
write.csv(hap_chr1_resistant.ihs$iHS,"ihs_resistant_chr1.csv")
write.csv(hap_chr1_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r1.csv")

hap_chr2_resilient.ihs<-ihh2ihs(hap_chr2_resilient.scan)
write.csv(hap_chr2_resilient.ihs$iHS,"ihs_resilient_chr2.csv")
write.csv(hap_chr2_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r2.csv")
hap_chr2_resistant.ihs<-ihh2ihs(hap_chr2_resistant.scan)
```

```
write.csv(hap_chr2_resistant.ihs$iHS,"ihs_resistant_chr2.csv")
write.csv(hap_chr2_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r2.csv")

hap_chr3_resilient.ihs<-ihh2ihs(hap_chr3_resilient.scan)
write.csv(hap_chr3_resilient.ihs$iHS,"ihs_resilient_chr3.csv")
write.csv(hap_chr3_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r3.csv")
hap_chr3_resistant.ihs<-ihh2ihs(hap_chr3_resistant.scan)
write.csv(hap_chr3_resistant.ihs$iHS,"ihs_resistant_chr3.csv")
write.csv(hap_chr3_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r3.csv")

hap_chr4_resilient.ihs<-ihh2ihs(hap_chr4_resilient.scan)
write.csv(hap_chr4_resilient.ihs$iHS,"ihs_resilient_chr4.csv")
write.csv(hap_chr4_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r4.csv")
hap_chr4_resistant.ihs<-ihh2ihs(hap_chr4_resistant.scan)
write.csv(hap_chr4_resistant.ihs$iHS,"ihs_resistant_chr4.csv")
write.csv(hap_chr4_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r4.csv")

hap_chr5_resilient.ihs<-ihh2ihs(hap_chr5_resilient.scan)
write.csv(hap_chr5_resilient.ihs$iHS,"ihs_resilient_chr5.csv")
write.csv(hap_chr5_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r5.csv")
hap_chr5_resistant.ihs<-ihh2ihs(hap_chr5_resistant.scan)
write.csv(hap_chr5_resistant.ihs$iHS,"ihs_resistant_chr5.csv")
write.csv(hap_chr5_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r5.csv")

hap_chr6_resilient.ihs<-ihh2ihs(hap_chr6_resilient.scan)
write.csv(hap_chr6_resilient.ihs$iHS,"ihs_resilient_chr6.csv")
write.csv(hap_chr6_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r6.csv")
hap_chr6_resistant.ihs<-ihh2ihs(hap_chr6_resistant.scan)
write.csv(hap_chr6_resistant.ihs$iHS,"ihs_resistant_chr6.csv")
write.csv(hap_chr6_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r6.csv")

hap_chr7_resilient.ihs<-ihh2ihs(hap_chr7_resilient.scan)
write.csv(hap_chr7_resilient.ihs$iHS,"ihs_resilient_chr7.csv")
write.csv(hap_chr7_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r7.csv")
hap_chr7_resistant.ihs<-ihh2ihs(hap_chr7_resistant.scan)
write.csv(hap_chr7_resistant.ihs$iHS,"ihs_resistant_chr7.csv")
write.csv(hap_chr7_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r7.csv")

hap_chr8_resilient.ihs<-ihh2ihs(hap_chr8_resilient.scan)
write.csv(hap_chr8_resilient.ihs$iHS,"ihs_resilient_chr8.csv")
write.csv(hap_chr8_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r8.csv")
hap_chr8_resistant.ihs<-ihh2ihs(hap_chr8_resistant.scan)
write.csv(hap_chr8_resistant.ihs$iHS,"ihs_resistant_chr8.csv")
```

```
write.csv(hap_chr8_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r8.csv")

hap_chr9_resilient.ihs<-ihh2ihs(hap_chr9_resilient.scan)
write.csv(hap_chr9_resilient.ihs$iHS,"ihs_resilient_chr9.csv")
write.csv(hap_chr9_resilient.ihs$frequency.class,"frequency_class_resilient_ch
r9.csv")
hap_chr9_resistant.ihs<-ihh2ihs(hap_chr9_resistant.scan)
write.csv(hap_chr9_resistant.ihs$iHS,"ihs_resistant_chr9.csv")
write.csv(hap_chr9_resistant.ihs$frequency.class,"frequency_class_resistant_ch
r9.csv")

hap_chr10_resilient.ihs<-ihh2ihs(hap_chr10_resilient.scan)
write.csv(hap_chr10_resilient.ihs$iHS,"ihs_resilient_chr10.csv")
write.csv(hap_chr10_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr10.csv")
hap_chr10_resistant.ihs<-ihh2ihs(hap_chr10_resistant.scan)
write.csv(hap_chr10_resistant.ihs$iHS,"ihs_resistant_chr10.csv")
write.csv(hap_chr10_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr10.csv")

hap_chr11_resilient.ihs<-ihh2ihs(hap_chr11_resilient.scan)
write.csv(hap_chr11_resilient.ihs$iHS,"ihs_resilient_chr11.csv")
write.csv(hap_chr11_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr11.csv")
hap_chr11_resistant.ihs<-ihh2ihs(hap_chr11_resistant.scan)
write.csv(hap_chr11_resistant.ihs$iHS,"ihs_resistant_chr11.csv")
write.csv(hap_chr11_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr11.csv")

hap_chr12_resilient.ihs<-ihh2ihs(hap_chr12_resilient.scan)
write.csv(hap_chr12_resilient.ihs$iHS,"ihs_resilient_chr12.csv")
write.csv(hap_chr12_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr12.csv")
hap_chr12_resistant.ihs<-ihh2ihs(hap_chr12_resistant.scan)
write.csv(hap_chr12_resistant.ihs$iHS,"ihs_resistant_chr12.csv")
write.csv(hap_chr12_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr12.csv")

hap_chr13_resilient.ihs<-ihh2ihs(hap_chr13_resilient.scan)
write.csv(hap_chr13_resilient.ihs$iHS,"ihs_resilient_chr13.csv")
write.csv(hap_chr13_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr13.csv")
hap_chr13_resistant.ihs<-ihh2ihs(hap_chr13_resistant.scan)
write.csv(hap_chr13_resistant.ihs$iHS,"ihs_resistant_chr13.csv")
write.csv(hap_chr13_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr13.csv")

hap_chr14_resilient.ihs<-ihh2ihs(hap_chr14_resilient.scan)
write.csv(hap_chr14_resilient.ihs$iHS,"ihs_resilient_chr14.csv")
write.csv(hap_chr14_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr14.csv")
hap_chr14_resistant.ihs<-ihh2ihs(hap_chr14_resistant.scan)
write.csv(hap_chr14_resistant.ihs$iHS,"ihs_resistant_chr14.csv")
write.csv(hap_chr14_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr14.csv")
```

```
hap_chr15_resilient.ihs<-ihh2ihs(hap_chr15_resilient.scan)
write.csv(hap_chr15_resilient.ihs$iHS,"ihs_resilient_chr15.csv")
write.csv(hap_chr15_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr15.csv")
hap_chr15_resistant.ihs<-ihh2ihs(hap_chr15_resistant.scan)
write.csv(hap_chr15_resistant.ihs$iHS,"ihs_resistant_chr15.csv")
write.csv(hap_chr15_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr15.csv")

hap_chr16_resilient.ihs<-ihh2ihs(hap_chr16_resilient.scan)
write.csv(hap_chr16_resilient.ihs$iHS,"ihs_resilient_chr16.csv")
write.csv(hap_chr16_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr16.csv")
hap_chr16_resistant.ihs<-ihh2ihs(hap_chr16_resistant.scan)
write.csv(hap_chr16_resistant.ihs$iHS,"ihs_resistant_chr16.csv")
write.csv(hap_chr16_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr16.csv")

hap_chr17_resilient.ihs<-ihh2ihs(hap_chr17_resilient.scan)
write.csv(hap_chr17_resilient.ihs$iHS,"ihs_resilient_chr17.csv")
write.csv(hap_chr17_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr17.csv")
hap_chr17_resistant.ihs<-ihh2ihs(hap_chr17_resistant.scan)
write.csv(hap_chr17_resistant.ihs$iHS,"ihs_resistant_chr17.csv")
write.csv(hap_chr17_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr17.csv")

hap_chr18_resilient.ihs<-ihh2ihs(hap_chr18_resilient.scan)
write.csv(hap_chr18_resilient.ihs$iHS,"ihs_resilient_chr18.csv")
write.csv(hap_chr18_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr18.csv")
hap_chr18_resistant.ihs<-ihh2ihs(hap_chr18_resistant.scan)
write.csv(hap_chr18_resistant.ihs$iHS,"ihs_resistant_chr18.csv")
write.csv(hap_chr18_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr18.csv")

hap_chr19_resilient.ihs<-ihh2ihs(hap_chr19_resilient.scan)
write.csv(hap_chr19_resilient.ihs$iHS,"ihs_resilient_chr19.csv")
write.csv(hap_chr19_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr19.csv")
hap_chr19_resistant.ihs<-ihh2ihs(hap_chr19_resistant.scan)
write.csv(hap_chr19_resistant.ihs$iHS,"ihs_resistant_chr19.csv")
write.csv(hap_chr19_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr19.csv")

hap_chr20_resilient.ihs<-ihh2ihs(hap_chr20_resilient.scan)
write.csv(hap_chr20_resilient.ihs$iHS,"ihs_resilient_chr20.csv")
write.csv(hap_chr20_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr20.csv")
hap_chr20_resistant.ihs<-ihh2ihs(hap_chr20_resistant.scan)
write.csv(hap_chr20_resistant.ihs$iHS,"ihs_resistant_chr20.csv")
write.csv(hap_chr20_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr20.csv")

hap_chr21_resilient.ihs<-ihh2ihs(hap_chr21_resilient.scan)
```

```
write.csv(hap_chr21_resilient.ihs$iHS,"ihs_resilient_chr21.csv")
write.csv(hap_chr21_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr21.csv")
hap_chr21_resistant.ihs<-ihh2ihs(hap_chr21_resistant.scan)
write.csv(hap_chr21_resistant.ihs$iHS,"ihs_resistant_chr21.csv")
write.csv(hap_chr21_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr21.csv")

hap_chr22_resilient.ihs<-ihh2ihs(hap_chr22_resilient.scan)
write.csv(hap_chr22_resilient.ihs$iHS,"ihs_resilient_chr22.csv")
write.csv(hap_chr22_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr22.csv")
hap_chr22_resistant.ihs<-ihh2ihs(hap_chr22_resistant.scan)
write.csv(hap_chr22_resistant.ihs$iHS,"ihs_resistant_chr22.csv")
write.csv(hap_chr22_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr22.csv")

hap_chr23_resilient.ihs<-ihh2ihs(hap_chr23_resilient.scan)
write.csv(hap_chr23_resilient.ihs$iHS,"ihs_resilient_chr23.csv")
write.csv(hap_chr23_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr23.csv")
hap_chr23_resistant.ihs<-ihh2ihs(hap_chr23_resistant.scan)
write.csv(hap_chr23_resistant.ihs$iHS,"ihs_resistant_chr23.csv")
write.csv(hap_chr23_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr23.csv")

hap_chr24_resilient.ihs<-ihh2ihs(hap_chr24_resilient.scan)
write.csv(hap_chr24_resilient.ihs$iHS,"ihs_resilient_chr24.csv")
write.csv(hap_chr24_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr24.csv")
hap_chr24_resistant.ihs<-ihh2ihs(hap_chr24_resistant.scan)
write.csv(hap_chr24_resistant.ihs$iHS,"ihs_resistant_chr24.csv")
write.csv(hap_chr24_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr24.csv")

hap_chr25_resilient.ihs<-ihh2ihs(hap_chr25_resilient.scan)
write.csv(hap_chr25_resilient.ihs$iHS,"ihs_resilient_chr25.csv")
write.csv(hap_chr25_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr25.csv")
hap_chr25_resistant.ihs<-ihh2ihs(hap_chr25_resistant.scan)
write.csv(hap_chr25_resistant.ihs$iHS,"ihs_resistant_chr25.csv")
write.csv(hap_chr25_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr25.csv")

hap_chr26_resilient.ihs<-ihh2ihs(hap_chr26_resilient.scan)
write.csv(hap_chr26_resilient.ihs$iHS,"ihs_resilient_chr26.csv")
write.csv(hap_chr26_resilient.ihs$frequency.class,"frequency_class_resilient_c
hr26.csv")
hap_chr26_resistant.ihs<-ihh2ihs(hap_chr26_resistant.scan)
write.csv(hap_chr26_resistant.ihs$iHS,"ihs_resistant_chr26.csv")
write.csv(hap_chr26_resistant.ihs$frequency.class,"frequency_class_resistant_c
hr26.csv")

#calculate rsb and xpehh
```

```
chr1.rsb<-
ies2rsb(hap_chr1_resilient.scan,hap_chr1_resistant.scan,"resilient_chr1","resi
stant_chr1")
write.csv(chr1.rsb,"rsb_chr1.csv")
chr1.xpehh<-
ies2xpehh(hap_chr1_resilient.scan,hap_chr1_resistant.scan,"resilient_chr1","re
sistant_chr1")
write.csv(chr1.xpehh,"xpehh_chr1.csv")

chr2.rsb<-
ies2rsb(hap_chr2_resilient.scan,hap_chr2_resistant.scan,"resilient_chr2","resi
stant_chr2")
write.csv(chr2.rsb,"rsb_chr2.csv")
chr2.xpehh<-
ies2xpehh(hap_chr2_resilient.scan,hap_chr2_resistant.scan,"resilient_chr2","re
sistant_chr2")
write.csv(chr2.xpehh,"xpehh_chr2.csv")

chr3.rsb<-
ies2rsb(hap_chr3_resilient.scan,hap_chr3_resistant.scan,"resilient_chr3","resi
stant_chr3")
write.csv(chr3.rsb,"rsb_chr3.csv")
chr3.xpehh<-
ies2xpehh(hap_chr3_resilient.scan,hap_chr3_resistant.scan,"resilient_chr3","re
sistant_chr3")
write.csv(chr3.xpehh,"xpehh_chr3.csv")

chr4.rsb<-
ies2rsb(hap_chr4_resilient.scan,hap_chr4_resistant.scan,"resilient_chr4","resi
stant_chr4")
write.csv(chr4.rsb,"rsb_chr4.csv")
chr4.xpehh<-
ies2xpehh(hap_chr4_resilient.scan,hap_chr4_resistant.scan,"resilient_chr4","re
sistant_chr4")
write.csv(chr4.xpehh,"xpehh_chr4.csv")

chr5.rsb<-
ies2rsb(hap_chr5_resilient.scan,hap_chr5_resistant.scan,"resilient_chr5","resi
stant_chr5")
write.csv(chr5.rsb,"rsb_chr5.csv")
chr5.xpehh<-
ies2xpehh(hap_chr5_resilient.scan,hap_chr5_resistant.scan,"resilient_chr5","re
sistant_chr5")
write.csv(chr5.xpehh,"xpehh_chr5.csv")

chr6.rsb<-
ies2rsb(hap_chr6_resilient.scan,hap_chr6_resistant.scan,"resilient_chr6","resi
stant_chr6")
write.csv(chr6.rsb,"rsb_chr6.csv")
chr6.xpehh<-
ies2xpehh(hap_chr6_resilient.scan,hap_chr6_resistant.scan,"resilient_chr6","re
sistant_chr6")
write.csv(chr6.xpehh,"xpehh_chr6.csv")
```

```
chr7.rsb<-
ies2rsb(hap_chr7_resilient.scan,hap_chr7_resistant.scan,"resilient_chr7","resi
stant_chr7")
write.csv(chr7.rsb,"rsb_chr7.csv")
chr7.xpehh<-
ies2xpehh(hap_chr7_resilient.scan,hap_chr7_resistant.scan,"resilient_chr7","re
sistant_chr7")
write.csv(chr7.xpehh,"xpehh_chr7.csv")

chr8.rsb<-
ies2rsb(hap_chr8_resilient.scan,hap_chr8_resistant.scan,"resilient_chr8","resi
stant_chr8")
write.csv(chr8.rsb,"rsb_chr8.csv")
chr8.xpehh<-
ies2xpehh(hap_chr8_resilient.scan,hap_chr8_resistant.scan,"resilient_chr8","re
sistant_chr8")
write.csv(chr8.xpehh,"xpehh_chr8.csv")

chr9.rsb<-
ies2rsb(hap_chr9_resilient.scan,hap_chr9_resistant.scan,"resilient_chr9","resi
stant_chr9")
write.csv(chr9.rsb,"rsb_chr9.csv")
chr9.xpehh<-
ies2xpehh(hap_chr9_resilient.scan,hap_chr9_resistant.scan,"resilient_chr9","re
sistant_chr9")
write.csv(chr9.xpehh,"xpehh_chr9.csv")

chr10.rsb<-
ies2rsb(hap_chr10_resilient.scan,hap_chr10_resistant.scan,"resilient_chr10","r
esistant_chr10")
write.csv(chr10.rsb,"rsb_chr10.csv")
chr10.xpehh<-
ies2xpehh(hap_chr10_resilient.scan,hap_chr10_resistant.scan,"resilient_chr10",
"resistant_chr10")
write.csv(chr10.xpehh,"xpehh_chr10.csv")

chr11.rsb<-
ies2rsb(hap_chr11_resilient.scan,hap_chr11_resistant.scan,"resilient_chr11","r
esistant_chr11")
write.csv(chr11.rsb,"rsb_chr11.csv")
chr11.xpehh<-
ies2xpehh(hap_chr11_resilient.scan,hap_chr11_resistant.scan,"resilient_chr11",
"resistant_chr11")
write.csv(chr11.xpehh,"xpehh_chr11.csv")

chr12.rsb<-
ies2rsb(hap_chr12_resilient.scan,hap_chr12_resistant.scan,"resilient_chr12","r
esistant_chr12")
write.csv(chr12.rsb,"rsb_chr12.csv")
chr12.xpehh<-
ies2xpehh(hap_chr12_resilient.scan,hap_chr12_resistant.scan,"resilient_chr12",
"resistant_chr12")
write.csv(chr12.xpehh,"xpehh_chr12.csv")
```

```
chr13.rsb<-
ies2rsb(hap_chr13_resilient.scan,hap_chr13_resistant.scan,"resilient_chr13","r
esistant_chr13")
write.csv(chr13.rsb,"rsb_chr13.csv")
chr13.xpehh<-
ies2xpehh(hap_chr13_resilient.scan,hap_chr13_resistant.scan,"resilient_chr13",
"resistant_chr13")
write.csv(chr13.xpehh,"xpehh_chr13.csv")

chr14.rsb<-
ies2rsb(hap_chr14_resilient.scan,hap_chr14_resistant.scan,"resilient_chr14","r
esistant_chr14")
write.csv(chr14.rsb,"rsb_chr14.csv")
chr14.xpehh<-
ies2xpehh(hap_chr14_resilient.scan,hap_chr14_resistant.scan,"resilient_chr14",
"resistant_chr14")
write.csv(chr14.xpehh,"xpehh_chr14.csv")

chr15.rsb<-
ies2rsb(hap_chr15_resilient.scan,hap_chr15_resistant.scan,"resilient_chr15","r
esistant_chr15")
write.csv(chr15.rsb,"rsb_chr15.csv")
chr15.xpehh<-
ies2xpehh(hap_chr15_resilient.scan,hap_chr15_resistant.scan,"resilient_chr15",
"resistant_chr15")
write.csv(chr15.xpehh,"xpehh_chr15.csv")

chr16.rsb<-
ies2rsb(hap_chr16_resilient.scan,hap_chr16_resistant.scan,"resilient_chr16","r
esistant_chr16")
write.csv(chr16.rsb,"rsb_chr16.csv")
chr16.xpehh<-
ies2xpehh(hap_chr16_resilient.scan,hap_chr16_resistant.scan,"resilient_chr16",
"resistant_chr16")
write.csv(chr16.xpehh,"xpehh_chr16.csv")

chr17.rsb<-
ies2rsb(hap_chr17_resilient.scan,hap_chr17_resistant.scan,"resilient_chr17","r
esistant_chr17")
write.csv(chr17.rsb,"rsb_chr17.csv")
chr17.xpehh<-
ies2xpehh(hap_chr17_resilient.scan,hap_chr17_resistant.scan,"resilient_chr17",
"resistant_chr17")
write.csv(chr17.xpehh,"xpehh_chr17.csv")

chr18.rsb<-
ies2rsb(hap_chr18_resilient.scan,hap_chr18_resistant.scan,"resilient_chr18","r
esistant_chr18")
write.csv(chr18.rsb,"rsb_chr18.csv")
chr18.xpehh<-
ies2xpehh(hap_chr18_resilient.scan,hap_chr18_resistant.scan,"resilient_chr18",
"resistant_chr18")
write.csv(chr18.xpehh,"xpehh_chr18.csv")
```

```
chr19.rsb<-
ies2rsb(hap_chr19_resilient.scan,hap_chr19_resistant.scan,"resilient_chr19","r
esistant_chr19")
write.csv(chr19.rsb,"rsb_chr19.csv")
chr19.xpehh<-
ies2xpehh(hap_chr19_resilient.scan,hap_chr19_resistant.scan,"resilient_chr19",
"resistant_chr19")
write.csv(chr19.xpehh,"xpehh_chr19.csv")

chr20.rsb<-
ies2rsb(hap_chr20_resilient.scan,hap_chr20_resistant.scan,"resilient_chr20","r
esistant_chr20")
write.csv(chr20.rsb,"rsb_chr20.csv")
chr20.xpehh<-
ies2xpehh(hap_chr20_resilient.scan,hap_chr20_resistant.scan,"resilient_chr20",
"resistant_chr20")
write.csv(chr20.xpehh,"xpehh_chr20.csv")

chr21.rsb<-
ies2rsb(hap_chr21_resilient.scan,hap_chr21_resistant.scan,"resilient_chr21","r
esistant_chr21")
write.csv(chr21.rsb,"rsb_chr21.csv")
chr21.xpehh<-
ies2xpehh(hap_chr21_resilient.scan,hap_chr21_resistant.scan,"resilient_chr21",
"resistant_chr21")
write.csv(chr21.xpehh,"xpehh_chr21.csv")

chr22.rsb<-
ies2rsb(hap_chr22_resilient.scan,hap_chr22_resistant.scan,"resilient_chr22","r
esistant_chr22")
write.csv(chr22.rsb,"rsb_chr22.csv")
chr22.xpehh<-
ies2xpehh(hap_chr22_resilient.scan,hap_chr22_resistant.scan,"resilient_chr22",
"resistant_chr22")
write.csv(chr22.xpehh,"xpehh_chr22.csv")

chr23.rsb<-
ies2rsb(hap_chr23_resilient.scan,hap_chr23_resistant.scan,"resilient_chr23","r
esistant_chr23")
write.csv(chr23.rsb,"rsb_chr23.csv")
chr23.xpehh<-
ies2xpehh(hap_chr23_resilient.scan,hap_chr23_resistant.scan,"resilient_chr23",
"resistant_chr23")
write.csv(chr23.xpehh,"xpehh_chr23.csv")

chr24.rsb<-
ies2rsb(hap_chr24_resilient.scan,hap_chr24_resistant.scan,"resilient_chr24","r
esistant_chr24")
write.csv(chr24.rsb,"rsb_chr24.csv")
chr24.xpehh<-
ies2xpehh(hap_chr24_resilient.scan,hap_chr24_resistant.scan,"resilient_chr24",
"resistant_chr24")
write.csv(chr24.xpehh,"xpehh_chr24.csv")
```

```
chr25.rsb<-
ies2rsb(hap_chr25_resilient.scan,hap_chr25_resistant.scan,"resilient_chr25","r
esistant_chr25")
write.csv(chr25.rsb,"rsb_chr25.csv")
chr25.xpehh<-
ies2xpehh(hap_chr25_resilient.scan,hap_chr25_resistant.scan,"resilient_chr25",
"resistant_chr25")
write.csv(chr25.xpehh,"xpehh_chr25.csv")

chr26.rsb<-
ies2rsb(hap_chr26_resilient.scan,hap_chr26_resistant.scan,"resilient_chr26","r
esistant_chr26")
write.csv(chr26.rsb,"rsb_chr26.csv")
chr26.xpehh<-
ies2xpehh(hap_chr26_resilient.scan,hap_chr26_resistant.scan,"resilient_chr26",
"resistant_chr26")
write.csv(chr26.xpehh,"xpehh_chr26.csv")
```