

University of Wrocław  
Faculty of Biotechnology

Ph.D. Thesis



# **TagSNP selection based on bovine microarray data**

by

**Adrian Drożdż**

Supervisor: dr hab. Joanna Szyda

Wrocław, 2011



*For my little daughter Susan, and for my wife Paulina...*



## Abstrakt

Wybór tagSNP na podstawie danych pochodzących z bydlęcej macierzy SNP.

Adrian Drożdż

Katedra Genetyki i Ogólnej Hodowli Zwierząt, Uniwersytet Przyrodniczy we Wrocławiu,  
ul. Koźuchowska 7, 51-631 Wrocław

Polimorfizmy pojedynczego nukleotydu (SNP) są źródłem największej zmienności w organizmach. Z uwagi na ich powszechność występowania w genomie, jak również równomierne rozłożenie zarówno w rejonach kodujących jak i nie kodujących, są niezastąpionym źródłem informacji o zmienności genetycznej. W ostatnich kilku latach okazało się, że za znaczną część zmienności odpowiadają polimorfizmy typu CNV (ang. copy-number variation), czyli zmienna ilość kopii fragmentów DNA. Jednak ilość, charakter i rozmieszczenie CNV jest zupełnie odmienne, przez co SNP nadal są niezastąpionym źródłem informacji o zmienności genetycznej w populacji. W ostatniej dekadzie SNP stały się niezwykle popularne jako markery genetyczne. Są one szeroko badane, a ich ilość w genomach sięga milionów. W bardzo szybkim czasie znalazły zastosowanie w takich dziedzinach jak medycyna, biotechnologia, farmacja, czy rolnictwo.

Ogromna ilość dostępnych SNP z jednej strony pozwala na bardzo precyzyjne scharakteryzowanie zmienności genetycznej lecz z drugiej strony utrudnia obróbkę danych i znacznie wydłuża czas analizy. Przy dużej liczbie dostępnych SNP występuje efekt powtarzalności informacji, ponieważ SNP leżące blisko siebie często niosą tę samą albo podobną informację. Problem ten można rozwiązać wykorzystując zjawisko nierównowagi sprzężeń (linkage disequilibrium, LD). Badając korelację pomiędzy SNP fizycznie leżącymi blisko siebie, można stwierdzić, jak często dziedziczą się one razem. Jeżeli pomiędzy SNP występuje wysoka korelacja, do dalszej analizy wystarczy wybrać tylko jeden SNP, który będzie identyfikował (tagował) mały fragment chromosomu. Otrzymujemy w ten sposób zestaw tzw. tagSNP. Zbiór ten jest charakterystyczny dla danej populacji.

W mojej pracy, wykorzystując dane obejmujące genotypy 54 001 SNP dla 1 228 buhajów rasy polskiej Holsztyńsko-Fryzyjskiej (HF), przeprowadzono charakterystykę nierównowagi sprzężeń oraz wybór tagSNP dla całej populacji oraz sześciu wybranych podgrup zwierząt. Do wyboru tagSNP zastosowano cztery różne zestawy parametrów określonych przez maksymalne LD i minimalną frekwencję rzadszego allelu (MAF). W celu sprawdzenia czy struktura genetyczna populacji wpływa na wybór tagSNP (tj. współczynnik pokrewieństwa i inbredu), spośród wszystkich 1 228 buhajów wyróżniono 6 podgrup liczących 450 osobników różniących się strukturą spokrewnienia.

Okazało się, że trzy chromosomy (23, 24 i 26) charakteryzują się wyższym LD i wykazują mniejszy spadek LD wraz z odległością, inaczej niż pozostałe chromosomy. Charakterystyka bloków LD wykazała istnienie 1 163

bloków, czyli regionów na chromosomie które z wysokim prawdopodobieństwem dziedziczą się razem. Liczba bloków jest zdecydowanie wyższa od dotychczas zidentyfikowanych w australijskiej i niemieckiej populacji buhajów rasy HF, dla których określono niewiele ponad 700 bloków. Nie zadowala jednak fakt, iż w każdej z analizowanych populacji bloki pokrywały bardzo małą część genomu. W populacji polskiej pokrycie genomu wynosiło 7,17%. Wynik ten był jednakże wyższy, niż uzyskany dla populacji australijskiej i niemieckiej, odpowiednio 2,27% oraz 4,67%. Różnice w wyborze tagSNP pomiędzy subpopulacjami okazały się małe. Dla wszystkich chromosomów średnio 93.88% tagSNP było identycznych pomiędzy grupami.

Analiza wykazała że różnice w strukturze genetycznej badanej populacji mają niewielki wpływ na wybór tagSNP a tym samym na uzyskany procent pokrycia genomu blokami. Przeprowadzona analiza pozwoli na skonstruowanie macierzy SNP dostosowanej do struktury genetycznej polskiej populacji bydła rasy Holsztyńsko-Fryzyjskiej.

Słowa kluczowe: polimorfizm pojedynczego nukleotydu, tagSNP, nierównowaga sprzężeń, blok LD.

## Abstract

TagSNP selection based on bovine microarray data.

Adrian Drożdż

Institute of Genetics and Animal Breeding, Wrocław University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wrocław, Poland

Single nucleotide polymorphisms (SNPs) are the source of the greatest variation in organisms. Due to their prevalence in the genome and their uniform distribution in both the coding and non-coding regions, they are an irreplaceable source of information about genetic variation. Over the past few years, it has become clear that much of the variation corresponds to copy number variation (CNV), i.e. a variation in the number of copies of DNA fragments. The number, nature and location of CNVs might be different but SNPs are still an indispensable source of information about genetic variation. In the last decade, SNPs have become extremely popular as genetic markers. They are widely studied and genomes have millions of them. In this short time, they have found application in such fields as medicine, biotechnology, pharmacy and agriculture.

While the huge number of SNPs allows for a very precise description of genetic variation, it impedes data processing and significantly increases analysis time. Many of the SNPs situated close to each other frequently carry the same or similar information. A large number of SNPs therefore leads to redundancy and repetition. This problem can be solved by using the phenomenon of linkage disequilibrium (LD). The frequency with which SNPs inherit together can be obtained by analysing the correlation between SNPs in physical proximity. A representative SNP can be chosen to tag a small chromosome fragment, provided there is a high correlation between SNPs. A set of tagSNPs can be created this way. This collection of tagSNPs is specific to the population.

In this paper, 54 001 SNP genotypes for 1 228 Polish Holstein-Friesian (HF) bulls were used to characterise linkage disequilibrium and to select tagSNPs for: (i) the entire population; and (ii) six selected subgroups of animals.

For tagSNP selection, four different sets of parameters, specified by the maximum LD and the minimum minor allele frequency (MAF), were used. In order to check whether the genetic structure of the population (kinship and inbreeding coefficients) affects the tagSNP selection, among all of 1 228 bulls, six sub-population were selected (450 individuals each), each having a different relationship structure.

Chromosomes 23, 24 and 26 exhibited higher LD and less LD decay with distance than the other chromosomes. LD block characterisation revealed 1 163 blocks, i.e. chromosome regions having a high probability of inheriting together. The number of blocks is much higher than previously reported

for the Australian and German HF populations, for whom a little over 700 blocks were identified. However, it is not satisfied that in each of the analysed populations genome coverage by the blocks was very low. Genome coverage was 7.17% in the Polish population. However, this result was higher than that obtained for the Australian and German populations, viz. 2.27%, and 4.67% respectively. Differences in tagSNP selection between sub-populations were small. On average, 93.88% of the tagSNPs were identical between groups for all the chromosomes.

The study showed that differences in the genetic structure of the analysed population have little influence on tagSNP selection and thus on the percentage of genome coverage by the blocks. The analysis will allow for the construction of a SNP microarray adjusted to the genetic structure of the Polish Holstein-Friesian cattle population.

Keywords: single nucleotide polymorphisms, tagSNP, linkage disequilibrium, LD block.

---

# Contents

<b>Contents</b>	<b>i</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Motivation . . . . .	5
1.2 Single Nucleotide Polymorphisms . . . . .	6
1.3 SNP Microarray . . . . .	7
1.4 Linkage Disequilibrium . . . . .	8
1.5 Linkage Disequilibrium SNP based tagging . . . . .	9
1.6 <i>Bos taurus</i> as a model organism . . . . .	11
1.7 Linkage disequilibrium and tagSNP selection in the bovine genome	12
1.7.1 Genome-wide linkage disequilibrium . . . . .	12
1.7.2 Non genome-wide linkage disequilibrium . . . . .	13
1.7.3 TagSNP selection . . . . .	13
1.8 The aim of the study . . . . .	13
<b>2 Material and Methods</b>	<b>15</b>
2.1 Animals . . . . .	15
2.2 SNP genotypes . . . . .	15
2.2.1 Genome-wide SNP distribution . . . . .	15
2.2.2 Distance between adjacent SNPs . . . . .	18
2.3 Minor Allele Frequency . . . . .	24
2.4 Linkage disequilibrium . . . . .	24
2.4.1 Measures of LD . . . . .	24
2.4.2 LD calculations for closely linked SNPs . . . . .	24
2.4.3 Determining the structure of LD blocks . . . . .	24
2.5 TagSNP selection . . . . .	25
2.5.1 TagSNP subsets . . . . .	26
<b>3 Results</b>	<b>29</b>

3.1	Linkage Disequilibrium . . . . .	29
3.2	TagSNP selection . . . . .	37
<b>4</b>	<b>Discussion</b>	<b>43</b>
	<b>Bibliography</b>	<b>47</b>
	<b>Appendices</b>	<b>53</b>
<b>A</b>	<b>LD measures</b>	<b>55</b>
<b>B</b>	<b>TagSNP selection</b>	<b>57</b>
	<b>List of Symbols and Abbreviations</b>	<b>61</b>
	<b>List of Figures</b>	<b>62</b>
	<b>List of Tables</b>	<b>64</b>

---

# Acknowledgements

This work was done at the Institute of Genetics and Animal Breeding at Wrocław University of Environmental and Life Sciences. It was financially supported by the Polish Ministry of Science and Higher Education, grant no. N N311 310436.

Most of all, I would like to thank my supervisor, Professor Joanna Szyda, for valuable advice, guidance, and support.

I would also like to thank Professor Stanisław Cebrat and “smORFland” for a lot of interesting discussions and advice.

Special thanks to Tomasz Suchocki, for a lot of things.

*There's real poetry in the real world. Science is the poetry of reality.*

*Richard Dawkins*



*Isn't it sad to go to your grave without ever wondering why you were born?  
Who, with such a thought, would not spring from bed, eager to resume discovering  
the world and rejoicing to be part of it?*

*Richard Dawkins*



---

# Introduction

## 1.1 Motivation

The last decade may be characterised by a vast increase in the amount of biological data available and, as such, was a highly significant one for genetics. Databases such as GenBank (*Benson et al., 2000*), 1000 Genome Project (*The 1000 Genomes Project Consortium, 2010*) and dbSNP (*Sherry et al., 2001*) have rapidly accumulated data, and are the source for many studies. This has correlated with the new disciplines of bioinformatics, genomics and proteomics. When the draft sequence of the human genome was published in *Nature* (*International Human Genome Sequencing Consortium, 2001*) on 15 February 2001 and in *Science* (*Venter and et al, 2001*) the following day, scientists realised it was a first milestone to understanding not only the human genome, but also those of other species. *Bos taurus* genome was sequenced by *The Bovine Genome Sequencing and Analysis Consortium et al. (2009)*. New kind of markers were discovered and started to play a dominant role in genetics (*Brookes, 1999*). Their enormous number in mammalian genomes gives scientists an unprecedented tool to study the foundation of genetic variation. The International HapMap Project was begun in 2002 (*The International HapMap Consortium, 2003*). The aim of this project is to identify a large number of small differences in genome between selected human populations while accounting for the sharing patterns (haplotypes) of these differences. Analysis on the scale of whole genomes was possible thanks to the continuously increasing computational power.

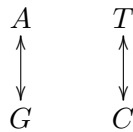
The rapid development of genetics and bioinformatics has influenced animal sciences, such as veterinary medicine and animal breeding. More and more genetic polymorphisms are being discovered for all domestic species, including cattle. Information about these polymorphisms is being put to practical use by many countries to predict additive genetic value of breeding animals, especially bulls and very young animals. This has given rise to a new kind of genetic analysis known as Genome-Wide Association Study (GWAS) (*McCarthy et al., 2008*).

Whole markers can be used to fingerprint an individual and associated with a given trait (or traits). It has been shown that estimating the breeding value of cattle this way can reduce costs by 97% (*Schaeffer, 2006*).

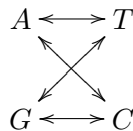
The results of these analyses play an important role in medicine through individual treatment of patients (e.g. *Allen et al., 2010; Mah and Chia, 2007*). Markers distributed over the entire bovine genome are utilised in this study. This is the first such analysis to describe the Polish Holstein-Friesian (HF) cattle population.

## 1.2 Single Nucleotide Polymorphisms

Genetic variation is the basis of differences between individuals. From all kinds of genetic variation **S**ingle **N**ucleotide **P**olymorphisms (SNP) play the main role as variation factors (*LaFramboise, 2009*) and are widely used as genetic markers in genetic studies (e.g. *Barendse et al., 2007*). From the point of view of molecular genetics, SNP is a single nucleotide difference between two DNA strains (point mutation). Table 1.1 shows an example of one SNP denoted as A/G. SNPs arise through transitions or transversions. A transition is a change of a purine into a purine (adenine, A into guanine, G), or a pyrimidine into a pyrimidine (thymine, T into cytosine, C):



A transversion is a change of a purine into a pyrimidine and vice versa:



Transversions change one DNA base into a completely different one in terms of shape and properties. They therefore have a much greater influence on DNA, and consequently on the resulting protein structure, than transitions. Insertions and deletions (indels) are less frequently encountered types of mutations of one or more DNA base pairs. These may cause the formation of SNPs (*Brookes, 1999*).

Table 1.1: Single Nucleotide Polymorphism in the DNA strands of two individuals. NB: only one chromosome is presented.

DNA of Individual A	5'	G	C	A	T	A	G	C	C	3'
DNA of Individual B	5'	G	C	G	T	A	G	C	C	3'
	3'	C	G	C	A	T	C	G	G	5'

Once a mutation’s frequency in the population exceeds a given threshold, it is regarded as SNP. A specific value for the threshold is unknown. For the human population, it is usually set to 1%. Therefore a punctual mutation which exceeds this frequency threshold is not called a mutation but a Single Nucleotide Polymorphism. Such polymorphism generally has two *alleles*<sup>1</sup>. The frequency of the rarer *allele* is denoted as **Minor Allele Frequency** (MAF).

Common SNP is defined as SNP shared by a group of animals (population, subpopulation). To avoid confusion and improve transparency in the description of the material in this work, SNP is considered as a marker placed on a microarray. The term “common SNP”, however, is not used. Table 1.2 illustrates the difference between mutation, SNP and “common SNP”.

Table 1.2: Frequencies of SNPs in different populations as a measure of SNP’s usability.

Candidate	Population X	Population Y	Population Z	Description
SNP1	0.001	0	0.0001	mutation or genotyping error
SNP2	0.2	0.23	0.499	valuable common SNP
SNP3	0.15	0.0001	0.001	SNP in Population X
SNP4	0.09	0.29	0.29	valuable common SNP
SNP5	0.01	0.009	0.015	common SNP or mutation/error

SNPs are widespread in genomes. Approximately 10 000 000 SNPs are predicted to exist in both the human and cattle genomes (*The Bovine HapMap Consortium, 2009; The International HapMap Consortium, 2003*). The International HapMap Project has so far verified 4 000 000 (*The International HapMap Consortium, 2003; Thorisson et al., 2005*). As many as 3 000 000 SNPs have been verified in cattle (*Affymetrix Inc.*).

### 1.3 SNP Microarray

Nowadays, the SNP microarray is a central tool in biomedical research (*LaFramboise, 2009*). In this study SNP genotypes were obtained using BeadArray<sup>TM</sup> technology developed by Illumina (*Oliphant et al., 2002*). This platform is commercially available as the BovineSNP50 microarray (fig. 1.1). Following *Matukumalli et al. (2009)*, one chip consists of 60 800 beads. From the total of 444 792 SNPs potentially available 58 336 were preselected for the BovineSNP50 assay (Table 1.3). The greedy algorithm was used for SNP selection (*Matukumalli et al., 2009*). In each iteration a new candidate SNP was placed into the largest remaining gap until the target density of 20 kbp was achieved or until no additional SNPs were available. SNPs with higher MAF were preferred. The SNPs originated from several sources (Table 1.3). Only publicly available SNPs were considered. Most of SNPs were discovered by *Matukumalli et al. (2009)* (44.3% of the assay). SNPs

<sup>1</sup>As there are four possible nucleotides, there are four possible *alleles* in the genomes. These, however, are rare and there is no point analysing them (*Brookes, 1999*)

from Bovine HapMap and Bovine Sequencing Projects comprised 50.6% of the assay. After that, the assay was tested on 576 animals (*Bovine* subfamily). 54 001 final markers were used on the microarray. Due to the genetic diversity between and within populations, not all of these markers are polymorphic in the Polish HF population.



Figure 1.1: Illumina BovineSNP50 microarray. Photography by Illumina Inc.

Table 1.3: Source of SNPs for the BovineSNP50 assay. Compilation based on *Matukumalli et al. (2009)*.

SNP source	Number of SNPs available for selection	Number of SNPs selected for the BovineSNP50 assay	% of BovineSNP50 assay
Draft <sup>1</sup>	235'725	10'244	17.6%
Interbreed <sup>2</sup>	73'127	6'035	10.3%
BAC <sup>3</sup>	36'387	1'526	2.6%
RRL <sup>4</sup>	65'180	25'833	44.3%
Bovine HapMap	29'853	13'236	22.7%
Parentage <sup>5</sup>	121	121	0.2%
Various <sup>6</sup>	4'399	1'341	2.3%
<b>Total</b>	<b>444'792</b>	<b>58'336</b>	<b>100%</b>

<sup>1,2</sup> SNPs from the bovine sequencing project (about 2.1 million SNPs)

<sup>3</sup> SNPs from aligning sequence traces from the Holstein Bacterial Artificial Chromosome to the Hereford genome assembly

<sup>4</sup> SNPs sequenced by *Matukumalli et al. (2009)*

<sup>5</sup> high quality SNPs

<sup>6</sup> all other sources

## 1.4 Linkage Disequilibrium

Linkage Disequilibrium (LD) explains a deviation of random *allele* association (*Boehnke, 2000*). Figure 1.2 visualises Linkage Disequilibrium in two populations and within 6 loci. The LD pattern is characteristic for each population (*Goldstein and Weale, 2001*), (Figure 1.2 and 1.3).

In livestock species, LD is mainly generated by selection and inbreeding, and decreased by crossing-over and mutations (*Kim and Kirkpatrick, 2009*). The

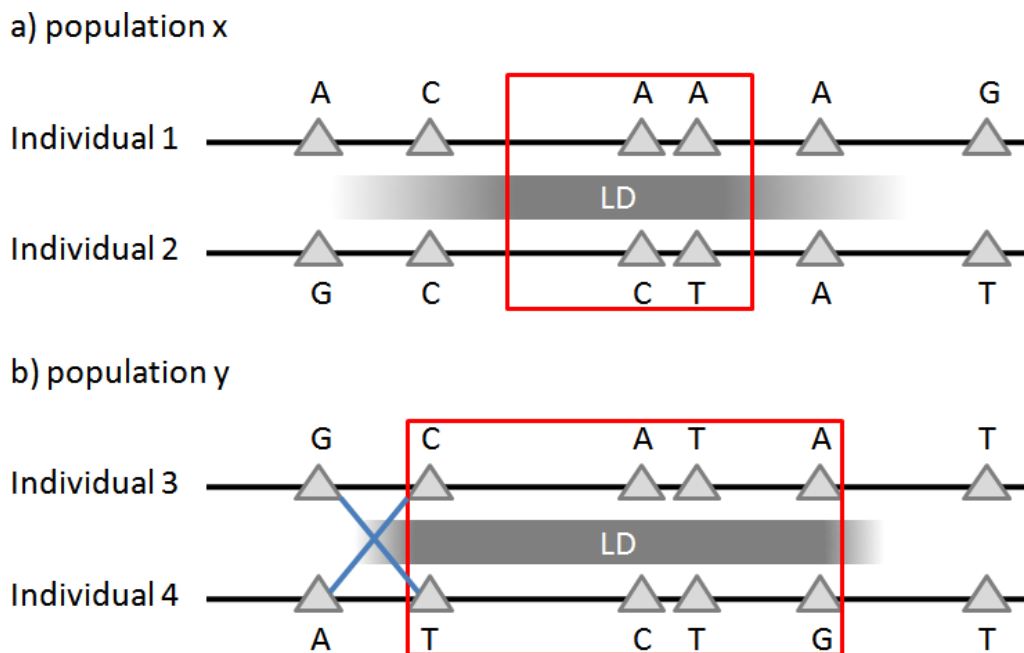


Figure 1.2: Linkage Disequilibrium in two populations. SNP alleles are denoted as DNA letters. For simplicity only 5'3' DNA is shown. The gray strip between the two pairs of individuals represents the LD strength - the darker the strip, the stronger the LD. In population  $x$ , LD is strongest (the strip is darkest) between loci  $A/C$  and  $A/T$ . It is highly probable that these loci are passed from generation to generation together (red rectangle). This region is called the LD block. In this high LD region one SNP may serve as tagSNP. While population  $y$  has a different LD pattern, the LD is strong at four loci ( $C/T$ ,  $A/C$ ,  $T/T$  and  $A/G$ ) and can be transmitted together. Crossing over (blue lines) occurs between loci  $G/A$  and  $C/T$ .

strength of LD varies across chromosomes.

## 1.5 Linkage Disequilibrium SNP based tagging

Genome-wide association studies (GWAS) uses a large number of SNPs. Not all of these markers are necessarily useful for the analysis. For this reason, one or more SNPs are represented by a so called tagSNP. This process is known as "tagging". As described by *Ding and Kullo (2007)* tagging methods can be classified as haplotype-dependent methods (e.g. the **haplotype tagging** method of *Johnson et al. (2001)*) and haplotype-independent methods (e.g. the **pairwise LD** method of *Carlson et al. (2004)*). The pairwise LD method is used in this study. A chosen tagSNP is both in high LD with other SNPs represented by this tagSNP and with low LD with other tagSNPs. Tagging reduces the number of SNPs needed for further analysis. This, in turn, cuts costs and computational time

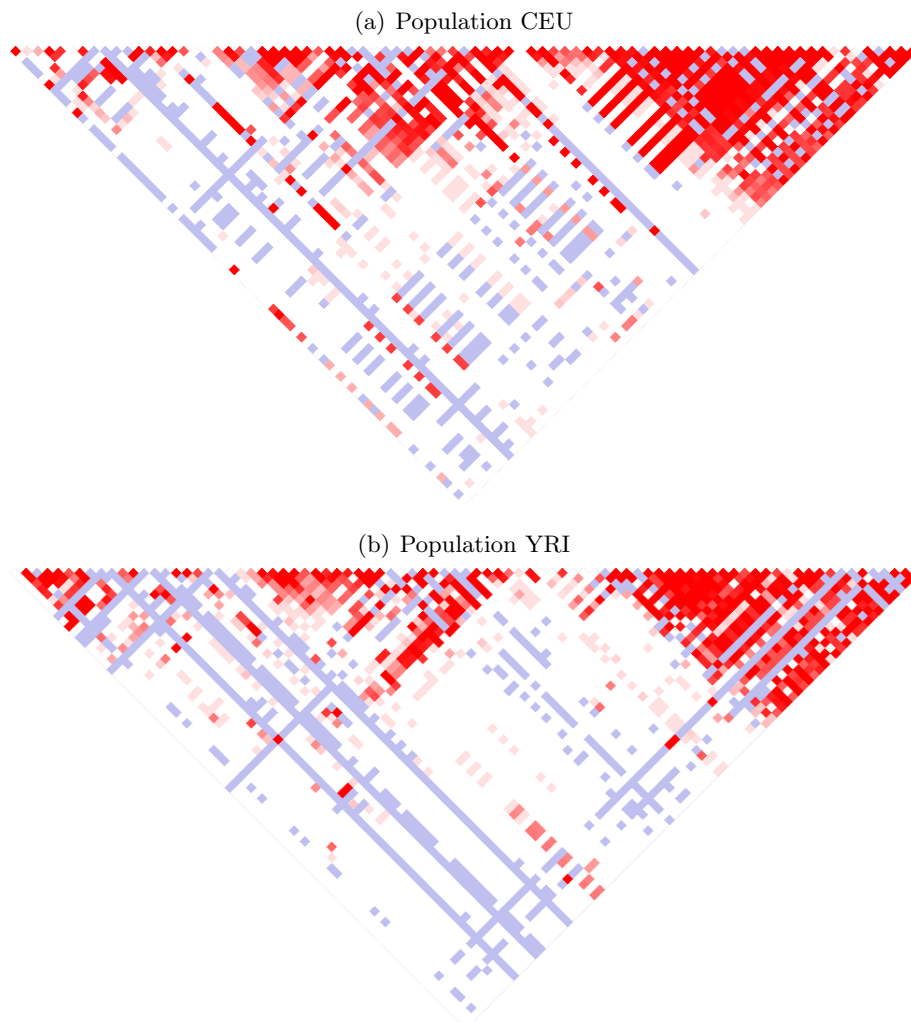


Figure 1.3: Linkage Disequilibrium in two human populations: CEU (a) and YRI (b). The 19th Chromosome from 2 200 to 2 400kb is presented. Each coloured square represents a LD calculated pairwise. Dark red and blue indicate the highest LD. Shades of pink and red indicate lower LD. White indicates a very low LD (*Barrett et al.*, 2004). The populations shown here have different LD patterns. Source: The International HapMap Project website (*Thorisson et al.*, 2005). A description of the populations can be found on page 61.

with a minimal decrease of dataset informativeness. This is because a tagSNP subset covers most of the variation in the genome (*LaFramboise, 2009*). SNP arrays such as BovineSNP50 were designed for population-wide use and are not constructed for any specific population. The main idea behind tagging the bovine SNP array is to adjust it to the population being analysed.

## 1.6 *Bos taurus* as a model organism

Cattle (*Bos primigenius taurus*<sup>2</sup>) and humped cattle (*Bos primigenius indicus*) are both auroch subspecies (*Bos primigenius primigenius*<sup>3</sup>). Cattle play a very important role in genetics, biology and agriculture (*Lewin, 2009*). The bovine genome was draft sequenced by *The Bovine Genome Sequencing and Analysis Consortium et al. (2009)*. It contains a minimum of 22 000 genes and spans 2.87 Gbp divided into 30 pairs of chromosomes (see Figure 1.4). The relatively long genetic distance to primates provides an opportunity to not only discover the cattle genome, but to understand human and other mammalian genome structure and evolution. For thousands of years, cattle have produced milk, meat, leather and have been used as labour force. In recent years, they have begun to serve as a commonly used model organism in genetics. With the increasing importance of SNPs in the study of humans and other species, scientists have started to discover more and more SNPs in the cattle genome (*Matukumalli et al., 2009; The Bovine Genome Sequencing and Analysis Consortium et al., 2009; The Bovine HapMap Consortium, 2009*). It is widely used in linkage disequilibrium studies as a representation of the *Bovinae* subfamily.



Figure 1.4: *Bos taurus* karyotype. All autosomes are acrocentric, i.e., their shorter arm is hard to observe. Source: <http://homepage.usask.ca/~schmutz/cowChroms.jpg>.

---

<sup>2</sup>*Bos taurus* in the rest of the text

<sup>3</sup>The species of the *Bos* genus can interbreed (e.g. yak, buntang and gaur)

## 1.7 Linkage disequilibrium and tagSNP selection in the bovine genome

As SNPs have started to play a dominant role in marker assisted analysis over the last few years, it was a matter of time before genome-wide SNP coverage became available. The most significant studies on the bovine genome are summarised below.

### 1.7.1 Genome-wide linkage disequilibrium

SNP based LD analyses for the cattle genome have been published in several papers. *Khatkar et al. (2007)* performed the first genome-wide linkage disequilibrium analysis for cattle. The authors used 15 036 SNPs, 10 410 of which were from the Affymetrix MegAllele GeneChip Bovine 10K SNP Array (*Affymetrix Inc., 2005; Hardenbol et al., 2005*)<sup>4</sup>, while remaining 4 626 SNPs were from other sources. 1 000 Australian Holstein-Friesian bulls were genotyped. At the same time, *McKay et al. (2007)* used 2 670 publicly available SNPs and 520 animals from the USA, Canada, Brazil, Japan and Belgium. The second study of *Khatkar et al.* was published in 2008. This study was strictly about LD in Australian HF cattle. The authors used the same set of SNPs (15 036) and more (1 546) bulls. In the same year *Sargolzaei et al.* analyzed 821 bulls from Canada and the USA. They used 5 564 SNPs genotyped using the Affymetrix 10K Array. The last study in 2008 was performed by *de Roos et al.* The authors analysed 3 987 individuals: 2 400 Dutch HF (3 072 SNPs); 379 Australian Angus (9 323 SNPs); 383 Australian HF (9 919 SNPs); 795 animals extracted from an  $F_2$  crossing experiment with New Zealand Jersey and New Zealand HF cattle (9 713 SNPs). SNPs for Dutch animals were selected *in silico* from the publicly available databases. SNPs for the Australian and New Zealand animals were genotyped using the Affymetrix 10K Array.

2009 was a very important year for genomic selection. *Illumina Inc.* developed a new microarray with 54 001 SNPs. *Matukumalli et al. (2009)* used 576 animals from different *Bovinae* species to validate a set of SNPs placed on an Illumina chip. In the same year, *Kim and Kirkpatrick* genotyped 200 American Holstein-Friesian sires using the Affymetrix 10K Array. From this array, 7 119 SNPs were chosen for a linkage disequilibrium analysis. Four papers on LD analysis using the Illumina BovineSNP50 (*Illumina Inc., 2009; Matukumalli et al., 2009*) were published in 2010. *Banos and Coffey* used 299 Greek Holstein-Friesian cows and 41 859 SNPs. *Bohmanova et al.* analyzed 887 North American Holstein-Friesian bulls using 38 590 SNPs. *Qanbari et al.* used 469 bulls and 341 bull dams belonging to the German HF population, of which 40 854 SNPs were considered.

---

<sup>4</sup>For more flexibility in the rest of the text, the full name of array will be shortened to “Affymetrix 10K Array”

### 1.7.2 Non genome-wide linkage disequilibrium

Four studies described linkage disequilibrium in selected regions only, i.e. not the whole bovine genome. *Gautier et al. (2007)* analysed the structure of linkage disequilibrium in 14 European and African cattle breeds. The authors used 1 536 SNPs discovered *in silico*. These SNPs were localized mostly on BTA1<sup>5</sup> (centromeric region), BTA3 (sequence similarity to the human genome) and BTA15 (telomeric region). 1 773 individuals were analysed. *Prasad et al. (2008)* used 440 Angus and Holstein-Friesian animals from the USA. They genotyped 1 001 evenly spaced SNPs on BTA19 and 506 on BTA29. *Marques et al. (2008)* used 331 Canadian Holstein-Friesian bulls and 137 Angus bulls from the USA. 843 SNPs were localized on BTA14 because of the importance of BTA14 in QTL analysis. *Villa-Angulo et al. (2009)* genotyped 32 826 SNPs from the Bovine HapMap Consortium database (*The Bovine Genome Sequencing and Analysis Consortium et al., 2009*) for 501 animals which were sampled from 19 cattle breeds (taurine and indicine) and two outgroups: Anoa (*Bubalus quarlesi* or *Bubalus depressicornis*) and Water Buffalo (*Bubalus bubalis*). Authors focused on dense SNP regions on BTA14, BTA25 and BTA6.

### 1.7.3 TagSNP selection

As for the whole-genome, the first and crucial genome-wide tagSNP analysis in the bovine genome was performed by *Khatkar et al. (2007)*. The Authors used both haplotype tagging and pairwise  $r^2$  tagging. The second genome-wide tagSNP analysis was published by *Qanbari et al. (2010)* when the Illumina BovineSNP50 became commercially available. The authors used the pairwise tagging method implemented in the Tagger software. There was a further analysis of pairwise tagSNP selection limited to BTA14 - the chromosome containing DGAT1 (a candidate gene for milk production traits) - conducted by *Marques et al. (2008)*.

## 1.8 The aim of the study

Until now, no linkage disequilibrium studies on Polish Holstein-Friesian population have been performed. Furthermore, no studies have compared the influence of reference population structure on LD and consequently on tagSNP selection. Therefore the aim of this study is to describe linkage disequilibrium in Polish Holstein-Friesian cattle and to select tagSNPs for this population. Moreover, the question as to whether population parameters (kinship and inbreeding coefficients) influence tagSNP selection is now being analysed for the first time.

---

<sup>5</sup>*Bos taurus* chromosome



---

# Material and Methods

## 2.1 Animals

The animals comprised 1 228 bulls from the HF dairy cattle population in Poland. The bulls were born between 1997 and 2003. Three of the bulls were unrelated with to any of the other bulls. The rest of the bulls were related to at least one other bull. The coefficient of relationship varied between 0 and 0.65 with a mean of  $0.03 \pm 0.04$  and a median of 0.02 (Figure 2.1) for the entire analysed population. The average coefficient of relationship for each bull was lower (Figure 2.2 a). It varied between 0 and 0.06 with a mean of  $0.03 \pm 0.08$  and a median of 0.03. The inbreeding coefficient of the analysed animals, calculated over ten generations, varied between 0 and 0.134 with a mean of  $0.09 \pm 0.0112$  and a median of 0.004 (Figure 2.2 b).

## 2.2 SNP genotypes

Each bull was genotyped with the BovineSNP50 BeadChip developed by Illumina Inc. (*Illumina Inc.*, 2009). Each chip consisted of 54 001 SNPs. Consequently, the total number of available records was 66 313 228.

### 2.2.1 Genome-wide SNP distribution

By default, the SNPs on the BovineSNP50 are positioned by BTAU 3.1 (*Matukumalli et al.*, 2009; *The Bovine Genome Sequencing and Analysis Consortium et al.*, 2009). 52 336 (96.9%) of the SNPs were mapped to chromosomes, while the remainder, consisting of 1 665 SNPs had no information regarding location and were therefore excluded from further analysis. Table 2.1 presents the number of SNPs for each *Bos taurus* chromosome. The number of SNPs varies from 3 343 for BTA1 to 747 for BTAX. Following *Zimin et al.* (2009), bovine chromosomes

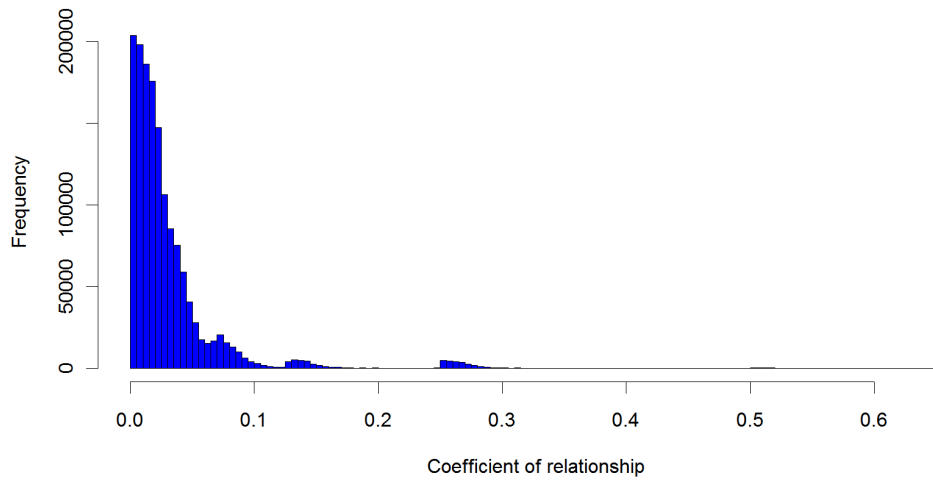


Figure 2.1: Coefficient of relationship.

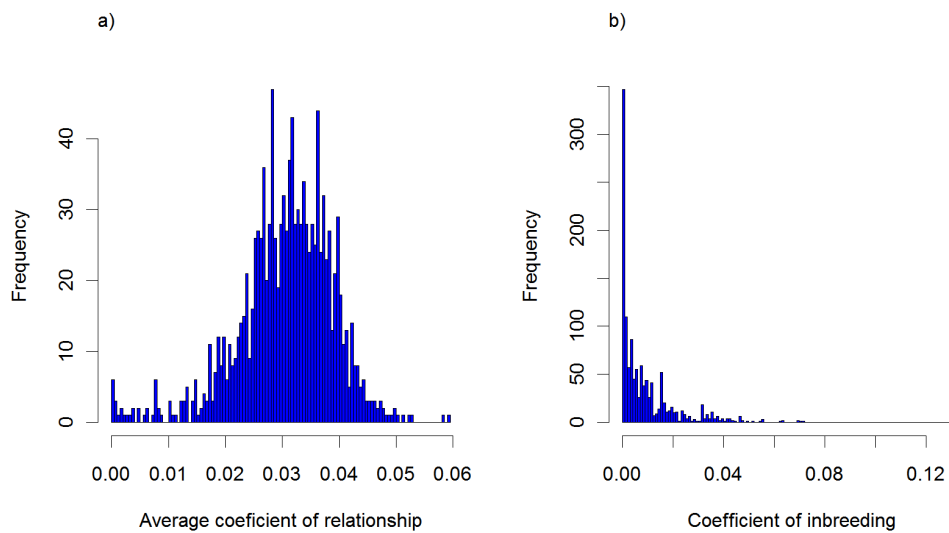


Figure 2.2: Average coefficient of relationship and inbreeding.

Table 2.1: Distribution of SNPs and genes across the genome.

Chromosome	Length [Mbp]	Number of SNPs	SNP density [SNP per Mbp]	Number of genes	Average distance [kbp]	Minimum distance [bp]	Maximum distance [kbp]
BTA1	161	3343	20.76	1040	48.16	1	631.95
BTA2	141	2764	19.60	1020	51.01	1	614.00
BTA3	128	2566	20.05	1535	49.88	1	784.00
BTA4	124	2541	20.49	870	48.80	1	783.91
BTA5	108	2181	17.31	630	57.77	147	1050.48
BTA6	126	2535	20.61	1399	48.52	49	599.69
BTA7	112	2294	20.48	1385	48.82	252	601.32
BTA8	123	2362	20.25	718	49.39	1	535.84
BTA9	110	2036	18.85	1014	53.05	49	729.11
BTA10	106	2179	20.56	1104	48.65	1	905.64
BTA11	117	2267	20.61	834	48.52	424	570.53
BTA12	85	1683	19.80	1258	50.51	238	568.86
BTA13	85	1802	21.45	433	46.61	1	488.00
BTA14	84	1722	21.26	797	47.04	1	352.35
BTA15	81	1688	19.86	502	50.36	1	609.40
BTA16	78	1606	20.59	622	48.57	179	671.82
BTA17	77	1585	20.58	631	48.58	332	725.53
BTA18	69	1351	20.47	582	48.85	772	633.90
BTA19	76	1378	21.20	401	47.17	993	543.54
BTA20	66	1564	20.58	1337	48.59	642	795.33
BTA21	65	1419	20.57	1269	48.63	1	586.50
BTA22	62	1299	20.95	571	47.73	1	360.64
BTA23	44	1083	20.43	763	48.94	305	442.97
BTA24	53	1294	19.91	910	50.23	32	417.55
BTA25	49	987	22.43	273	44.58	1915	543.93
BTA26	65	1086	20.88	358	47.88	1	286.63
BTA27	46	977	19.94	329	50.15	4669	1677.82
BTA28	52	942	20.48	685	48.83	1	470.56
BTA29	52	1048	20.15	437	49.62	1	734.85
BTAX	89	747	8.39	793	119.14	1168	999.08

differ in length starting from 44 Mbp for BTA23 to 161 Mbp for BTA1. The number of genes ranges from 273 in BTA25 to 1 535 in BTA3.

Figure 2.3 shows the scatter plots of SNPs, genes and chromosome length. BTAX outliers, and was therefore excluded from the comparison. Each scatter plot shows the regression line and the correlation coefficient. Comparing the amount of genotyped SNPs to base pairs per chromosome, one can assume that, except for BTAX (not presented on plots), the number of genotyped SNPs is proportional to the number of base pairs on each chromosome. The number of genes per chromosome is not proportional to the number of genotyped SNPs. For the BTA1 (161 Mbp, 3 343 SNPs, 1 040 genes), which is more than 2.4 times larger than the BTA20 (66 Mbp, 1 564 SNPs, 1 337 genes), has almost 300 fewer genes. There is a strong correlation between chromosome length and the number of SNPs ( $r = 0.99$ ), but a weak correlation between the number of genes and chromosome length ( $r = 0.52$ ), and the number of SNPs and the number of genes ( $r = 0.50$ ).

SNPs are almost evenly distributed across chromosomes as is shown on Figures 2.4 and 2.5. Only BTAX has considerably fewer SNPs than the other chro-

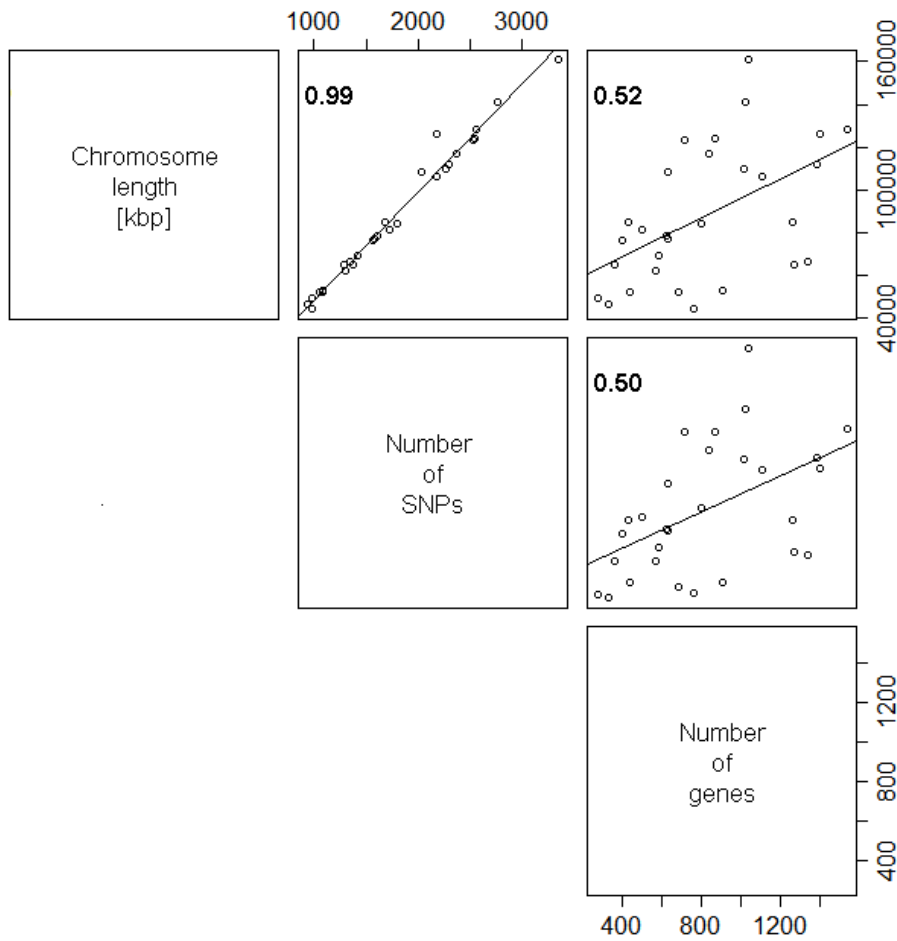


Figure 2.3: Scatter plots of chromosome length, the number of genes and SNPs.

mosomes. To measure the SNP density and distance between adjacent SNPs, the length of each chromosome was obtained from the more recent BTAU 4.1 (*The Bovine Genome Sequencing and Analysis Consortium et al., 2009*). On average, there were 20 SNPs per mega base pair. The lowest density was on BTAX (8.4 SNP/Mbp) and the highest on BTA25 (22.4 SNP/Mbp, Table 2.1).

### 2.2.2 Distance between adjacent SNPs

On average, the distance between two adjacent SNPs was  $51\,486 \pm 12\,969$  base pairs. There was little variation in average SNP spacing among autosomes (Table 2.1), with only BTAX having a markedly longer (119 143 bp) distance between adjacent SNPs. This measure is independent of chromosome length and is thus suitable for describing SNP coverage of each chromosome. Figures 2.6 and 2.7

show the distances between adjacent SNPs for each chromosome. The red line is the average distance between adjacent SNPs. The blue line is the target density for the greedy algorithm (Refer Section 1.3 on page 7). The algorithm sometimes allows shorter distances (*Matukumalli et al., 2009*). These are represented by points below the blue line. Some chromosomes have huge gaps, spanning close to 1 Mbp (e.g. BTA5). SNPs were rather evenly distributed along the genome, with only a few long gaps. Table 2.1 shows the minimum and maximum distance between SNPs for each chromosome. The smallest distance between two SNPs was 1 bp. BTA27 has the biggest gap with 1.68 Mbp. Furthermore, BTA27 has the largest minimum distance between SNPs (4 669 bp). The distance between adjacent SNPs is shorter than previously reported for the BovineSNP50 array (*Qanbari et al. (2010)* had 62.27 kbp, *Banos and Coffey (2010)* 57 kbp, *Bohmanova et al. (2010)* 66 kbp). Also *Bohmanova et al.* have longer average distance between adjacent SNPs on BTAX (215 kbp compared to 119 kbp in this study). In addition, SNP density, measured as SNP per 1 Mbp, was higher than reported by *Banos and Coffey (2010)* (17.5 SNP/Mbp).

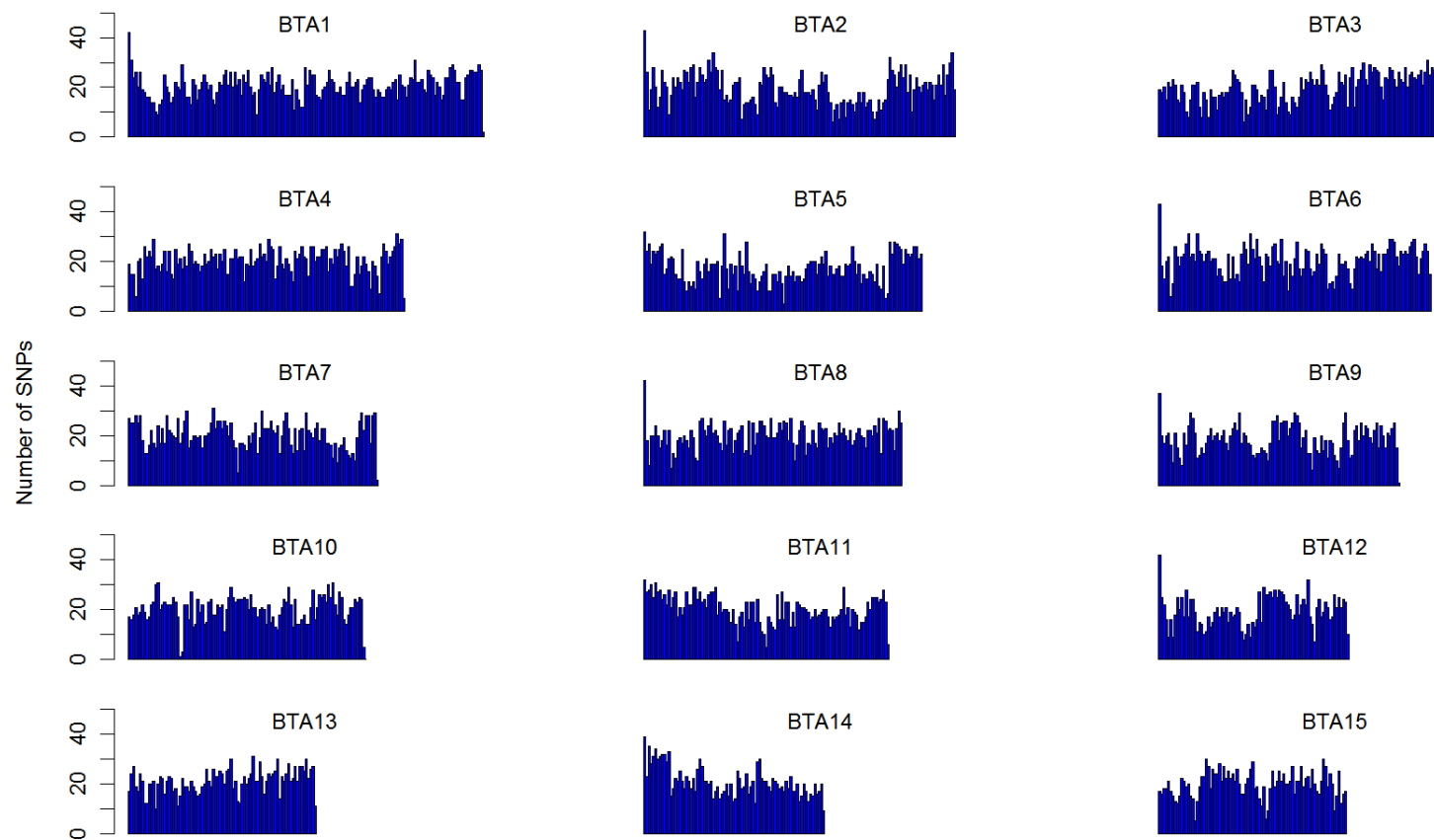


Figure 2.4: SNP distribution.

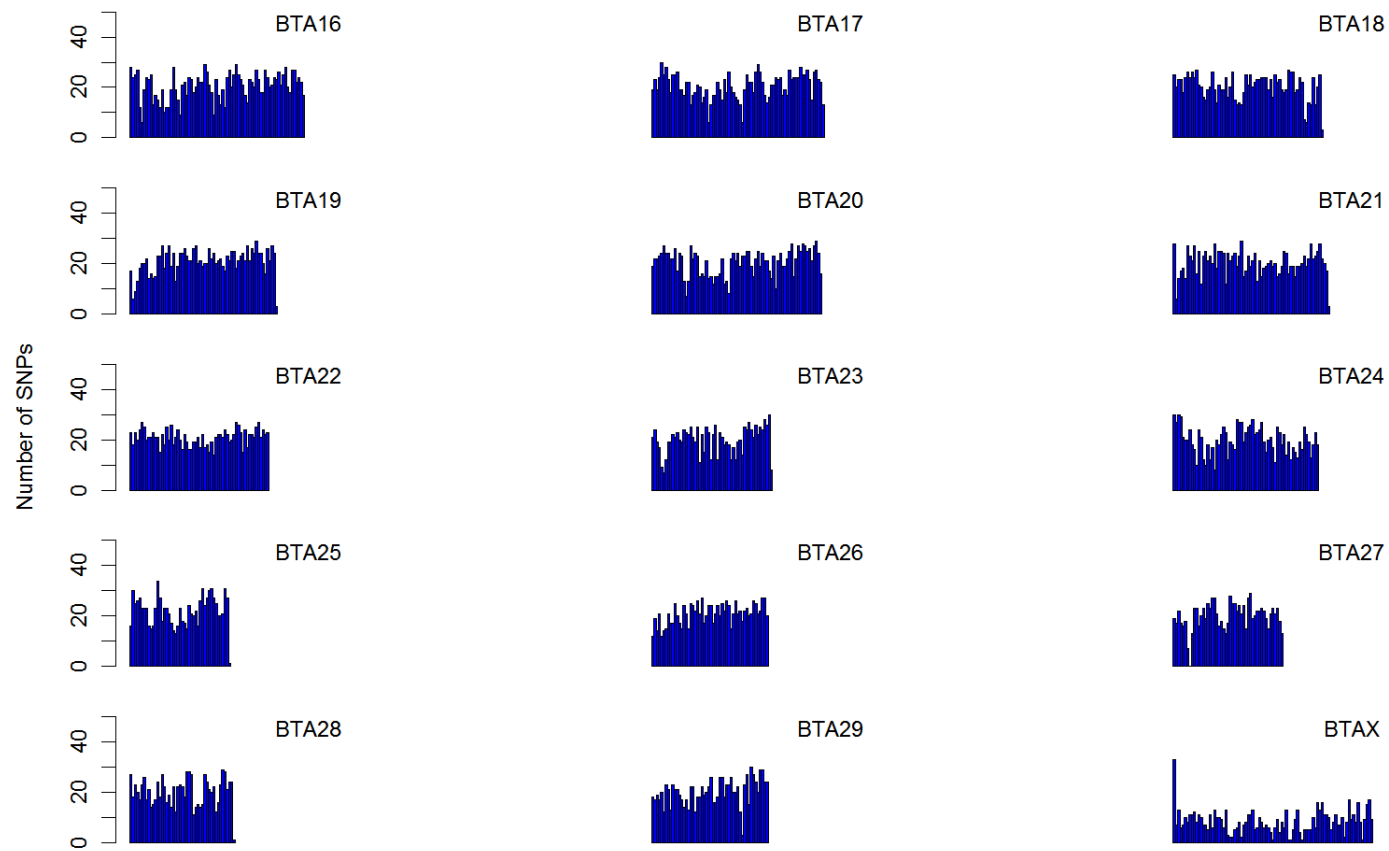


Figure 2.5: SNP distribution; continuation.

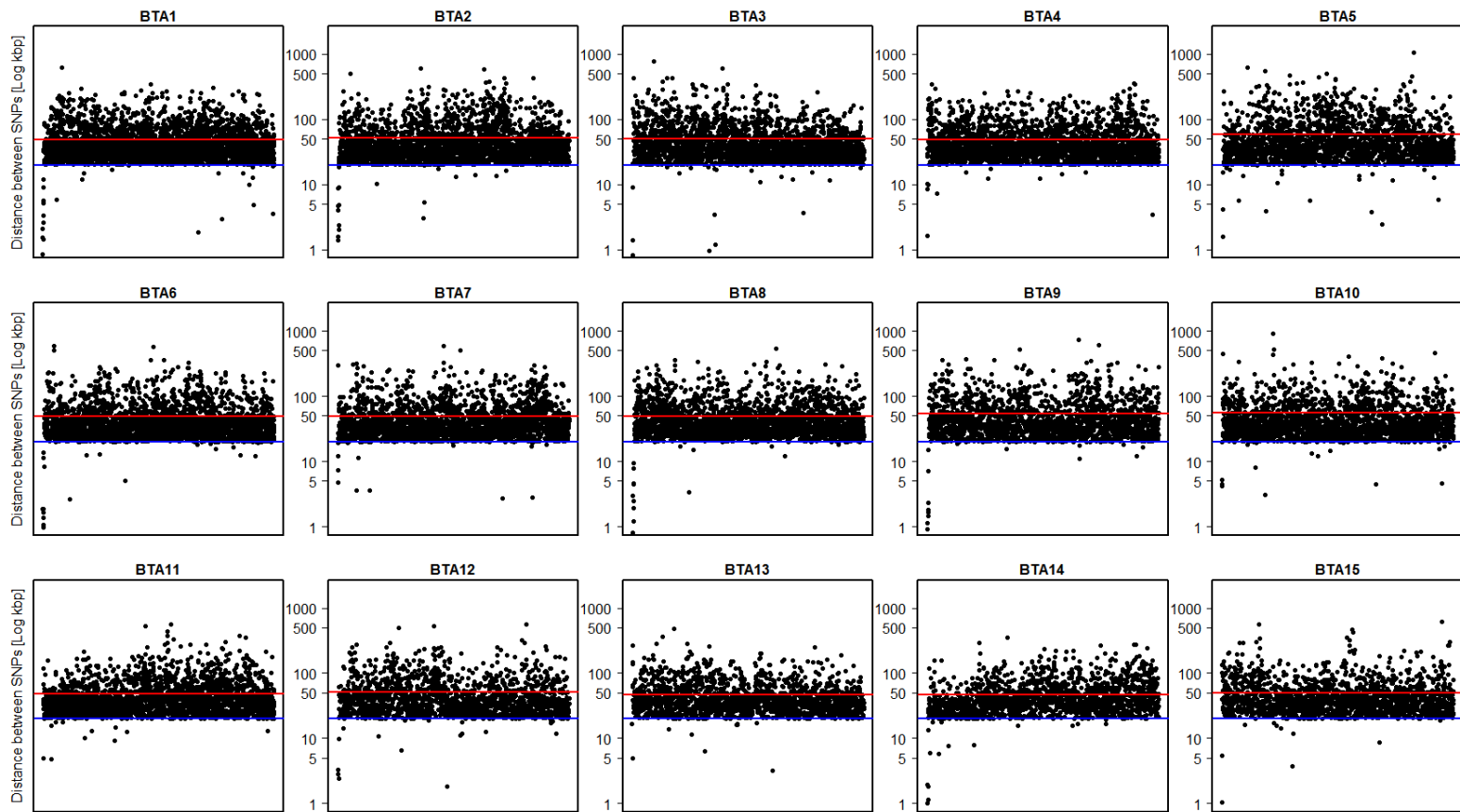


Figure 2.6: Distances between SPNs. The red line is an average distance between SNPs for each chromosome, while the blue line is a target value for the greedy algorithm.

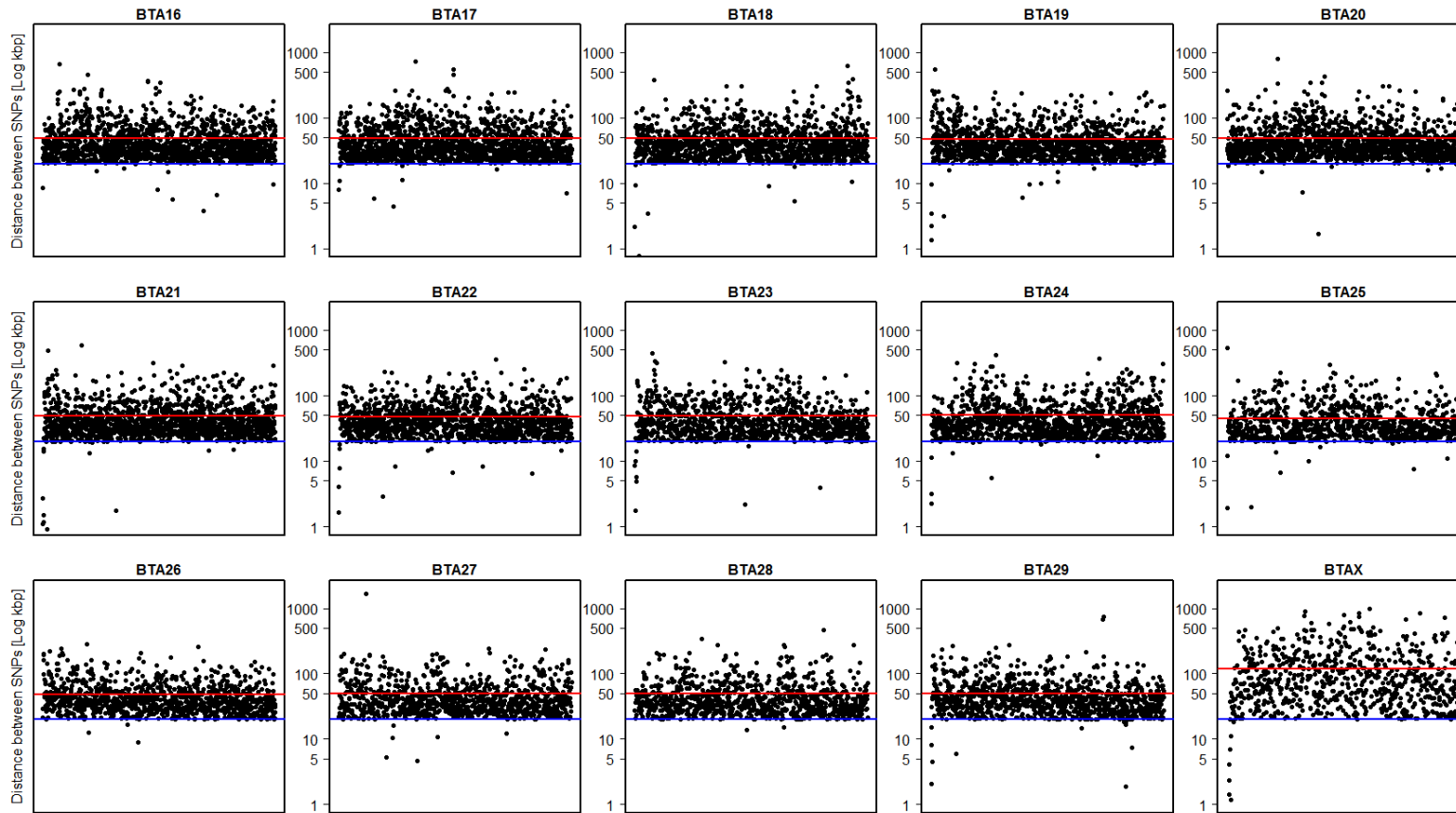


Figure 2.7: Distances between SPNs; continuation. The red line is an average distance between SNPs for each chromosome, while the blue line is a target value for the greedy algorithm.

## 2.3 Minor Allele Frequency

Minor allele frequency (MAF) varies from 0 to 0.5. Figure 2.8 presents the distribution of MAF. Those SNPs with the lowest MAF are overrepresented while the rest of the classes are represented uniformly. The average MAF for all chromosomes was  $0.227 \pm 0.0076$  with a median of 0.228. The mean MAF for each chromosome varied from 0.213 for BTAX to 0.245 for BTA25. The values for all SNPs calculated separately for each chromosome are presented in Figure 2.9. There is little variation between chromosomes.

## 2.4 Linkage disequilibrium

### 2.4.1 Measures of LD

Pairwise LD coefficients were quantified using  $D'$  and  $r^2$  statistics. Following *Lewontin and Kojima (1960)*,  $D'$  is defined as:

$$D' = \begin{cases} \frac{D^2}{D_{max}}, & \text{for } D \geq 0 \\ \frac{D^2}{D_{min}}, & \text{for } D < 0 \end{cases} \quad (2.1)$$

Where  $D$  is the deviation from the Hardy-Weinberg Equilibrium,  $D_{max}$  the smaller of  $p_1q_2$  and  $p_2q_1$ , and  $D_{min}$  is the larger of  $-p_1q_1$  and  $-p_2q_2$  (appendix A).

The pairwise  $r^2$  statistic was used as an alternative measure of LD (*Carlson et al., 2004; Devlin and Risch, 1995; Hill and Weir, 1994*):

$$r^2 = \frac{D^2}{p_1q_1p_2q_2}, \quad (2.2)$$

### 2.4.2 LD calculations for closely linked SNPs

PLINK software (*Purcell et al., 2007*) was used to calculate the pairwise LD for all linked SNP. For each SNP, calculations are made with neighbouring SNPs (maximum ten SNPs apart) within a 1 Mbp window.

### 2.4.3 Determining the structure of LD blocks

Haploview (*Barrett et al., 2004*) was used for LD block selection. Following (*Gabriel et al., 2002*), SNP pairs are in “strong LD” if the one-sided upper 95% confidence bound on  $D'$  is greater than 0.98 (no historical recombination) and the lower bound is above 0.7. On the other hand, “strong evidence for historical recombination” is if the upper confidence bound on  $D'$  is less than 0.9. SNPs with

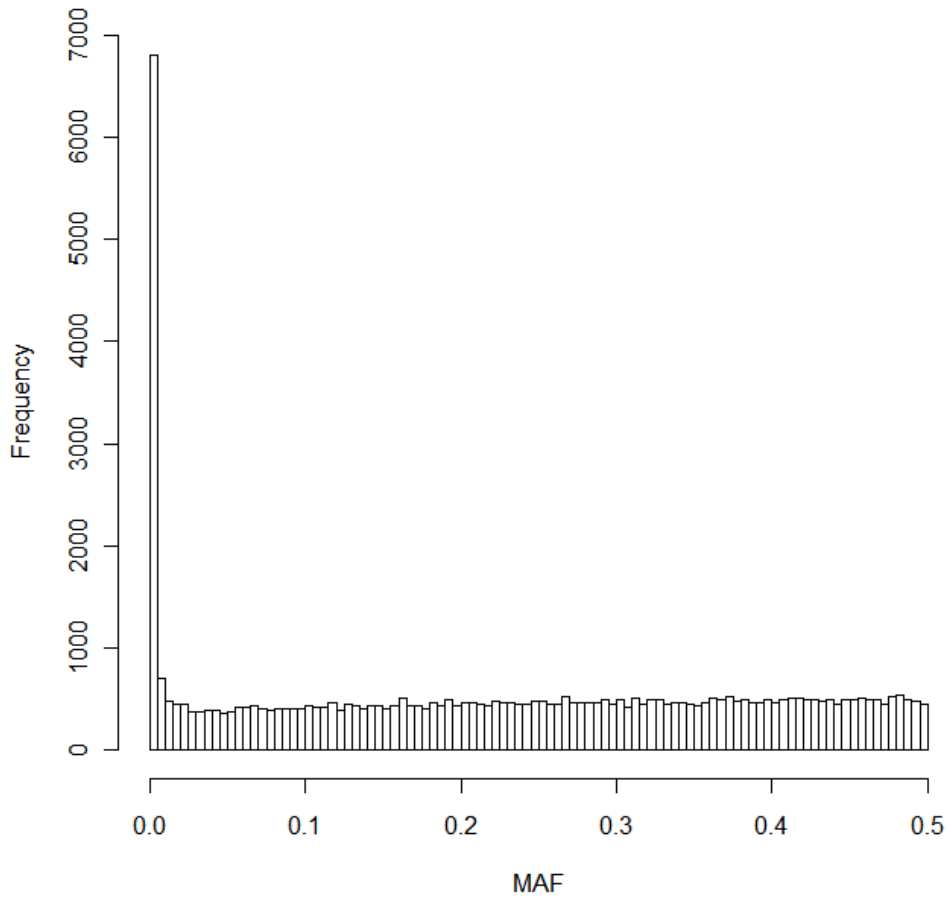


Figure 2.8: Genome-wide Distribution of MAF.

$MAF > 0.05$ , which belongs to these two categories, are considered to be informative. The confidence bounds on  $D'$ , rather than the point estimates of  $D'$  are used because the  $D'$  values are known to fluctuate upward when a small number of samples or rare *alleles* are examined. The confidence limits were determined by calculating the probability of the observed data for all possible values of  $D'$ , from which an overall probability distribution was determined. The proportion of informative pairs in strong LD in the block must be at least 95%.

## 2.5 TagSNP selection

The Tagger program (*de Bakker, 2009; de Bakker et al., 2005*) is implemented in Haploview (*Barrett, 2009; Barrett et al., 2004*) and used to select tagSNP

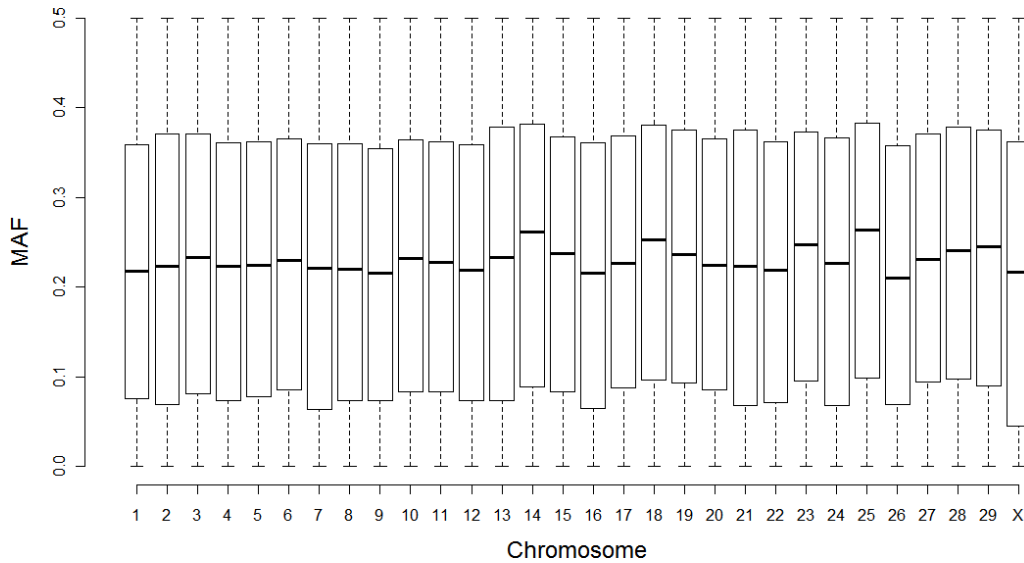


Figure 2.9: MAF on each chromosome.

subsets for each chromosome. TagSNPs were selected by the greedy pairwise tagging algorithm as described by *Carlson et al. (2004)*. Initially  $r^2$  is calculated for each SNP that exceeded the given MAF threshold. Once this step has been performed, each SNP whose maximum number of other SNPs exceeded the given  $r^2$  threshold is named as a bin. Since not all pairs in the bin exceed the  $r^2$  threshold, all pairwise  $r^2$  have to be re-evaluated in the bin in such a way that any SNP exceeding the  $r^2$  threshold with all other SNPs in the bin is selected as a tagSNP. There can therefore be more than one tagSNP per bin. SNPs that do not exceed the  $r^2$  threshold with other SNPs are named singletons. The fewer singletons there are, the fewer tagSNPs obtained.

### 2.5.1 TagSNP subsets

First, four subsets of tagSNPs were obtained with different  $r^2$  and MAF thresholds: (i) M5R8 subset with  $\text{MAF} \geq 0.05$  and  $r^2 \geq 0.8$ ; (ii) M5R5 subset with  $\text{MAF} \geq 0.05$  and  $r^2 \geq 0.5$ ; (iii) M1R8 subset with  $\text{MAF} \geq 0.01$  and  $r^2 \geq 0.8$ ; (iv) M1R5 subset with  $\text{MAF} \geq 0.01$  and  $r^2 \geq 0.5$ .

Furthermore, to investigate whether tagging depends on population parameters, different groups of animals were selected. Each such group consisted of 450 individuals. The first group (gr1) comprised animals having the lowest average coefficient of relationship (between 0 and 0.028). The second group (gr2) comprised animals having the highest average coefficient of relationship (between 0.034 and 0.046). The third group (gr3) comprised animals having the lowest coefficient of inbreeding (between 0 and 0.002197). The fourth group (gr4) comprised animals having the highest coefficient of inbreeding (between 0.007935 and 0.133789).

Table 2.2: The description of animal subsets.

Group	Description	Animals
gr1	lowest average coefficient of relationship	450
gr2	highest average coefficient of relationship	450
gr3	lowest coefficient of inbreeding	450
gr4	highest coefficient of inbreeding	450
gr5	first random sample	450
gr6	second random sample	450
gr7	all bulls (reference)	1228

Two random groups (gr5 and gr6) were also created from all the 1 228 available individuals. The last group (gr7) was a reference comprising all the available bulls. Finally, the seven groups (Table 2.2) were tagged based on M1R8.

The % of common individuals between groups is summarised in Table 2.3. Groups gr1, gr2, gr3 and gr4 contain no common individuals. The largest number of common animals in the other group comparisons are between gr1 and gr3 and between gr2 and gr4.

Table 2.3: The percentage of common individuals between groups.

<b>gr1</b>	0	56.67	20.67	38.00	33.11
	<b>gr2</b>	20.44	49.56	36.89	39.56
		<b>gr3</b>	0	40.00	32.44
			<b>gr4</b>	37.33	39.33
				<b>gr5</b>	39.56
					<b>gr6</b>



---

## Results

### 3.1 Linkage Disequilibrium

In this study linkage disequilibrium is presented in three ways. First,  $r^2$  between all linked SNPs is shown across each chromosome to indicate the different LD pattern in the chromosome regions. The mean value of  $r^2$  for each chromosome ranges from 0.07 for BTAX to 0.22 for BTA26. The distribution of  $r^2$  pairs across all chromosomes is presented in Figures 3.1 and 3.2. The red line is the  $r^2$  mean for each chromosome. BTA23, BTA24 and BTA26 have a significantly higher LD mean. The difference in LD pattern between chromosomes and the lack of  $r^2$  pairs can be observed in some regions. Moreover, while most of the  $r^2$  pairs are below 0.2, many pairs have  $r^2 = 1$ .

Second, the distance between  $r^2$  pairs is presented to show LD decay as a function of physical distance. Figures 3.3 and 3.4 show the distance (maximum 1 Mbp) between each  $r^2$  pair (maximum 10 SNPs apart) calculated for each chromosome. The red line is the  $r^2$  mean for each chromosome. Besides BTA23, BTA24 and BTA26, the decay of LD is clearly visible. As the distance between pairs increases, the LD rapidly decreases. Even for BTAX, it is quite easy to determine LD decay, although the density of SNPs and pairs for BTAX is low. BTA23, BTA24 and BTA26 have much more distant pairs in high LD.

Third, LD blocks are used to indicate the pieces of DNA which are inherited together with high probability. LD blocks are selected on each chromosome, but on BTAX it is not possible to determine blocks. The statistics summarising the attributes of blocks are listed in Table 3.1. The total number of discovered blocks is 1 163. The lowest number (8) is for BTA27 and BTA28 and the highest (97) is for BTA1. On average, the block length is 170.8 kbp, but for chromosomes taken separately it varies from 123.9 kbp for BTA22 to 199.3 kbp for BTA5. Moreover, the minimum block size is 0.5 kbp and the maximum is 499 kbp. The total length of blocks per chromosome varies from 17 357.5 kbp for BTA1 to 1 160.5 kbp for BTA28. Chromosome coverage (the percentage of total block

Table 3.1: Summary of LD blocks.

Chromosome	Total length [kbp]	Number	Average length [kbp]	Chromosome coverage [%]
BTA1	17357.5	97	178.94	10.78
BTA2	11832.0	71	166.65	8.39
BTA3	13305.5	78	170.58	10.39
BTA4	10656.5	60	177.61	8.59
BTA5	9567.5	48	199.32	7.59
BTA6	13598.0	72	188.86	11.05
BTA7	9854.0	53	185.92	8.80
BTA8	12332.5	66	186.86	10.54
BTA9	6626.5	34	194.90	6.13
BTA10	8933.0	63	141.79	8.43
BTA11	7759.0	50	155.18	7.05
BTA12	6127.0	39	157.10	7.21
BTA13	7894.5	46	171.62	9.40
BTA14	7874.0	40	196.85	9.72
BTA15	5631.5	36	156.43	6.62
BTA16	8309.0	46	180.63	10.65
BTA17	4653.5	29	160.46	6.04
BTA18	2889.0	19	152.05	4.38
BTA19	5146.0	29	177.45	7.92
BTA20	5781.0	31	186.48	7.61
BTA21	4371.5	28	156.12	6.34
BTA22	3718.0	30	123.93	6.00
BTA23	1711.0	12	142.58	3.23
BTA24	2346.0	18	130.33	3.61
BTA25	2836.0	19	149.26	6.44
BTA26	2743.0	17	161.35	5.27
BTA27	1448.0	8	181.00	2.95
BTA28	1160.5	8	145.06	2.52
BTA29	2236.5	16	139.78	4.30
BTAX	0	0	0	0

length per chromosome) varies from 11.5% for BTA6 to 2.52% for BTA28. The average coverage is 7.17%. In addition, Figures 3.5 and 3.6 show the frequency of block length in each chromosome, except for BTAX. The number of blocks and block length vary for all chromosomes. For example, BTA18 has a narrow range of block length with the shortest block spanning 87 kbp and the longest block spanning 242 kbp, while most chromosomes have a wide range of block length.

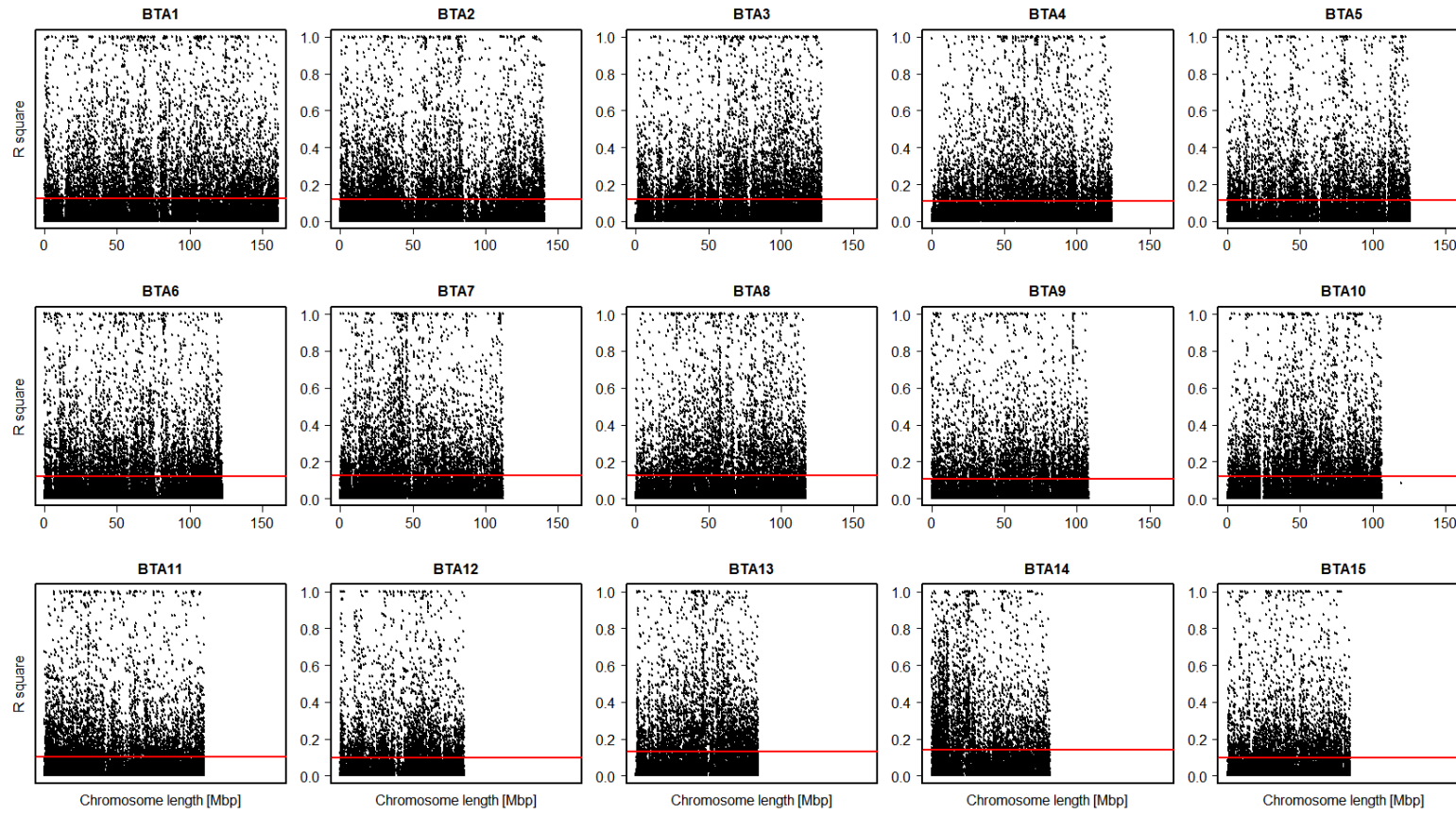


Figure 3.1: Linkage disequilibrium across chromosomes. Red line is  $r^2$  mean for each chromosome.

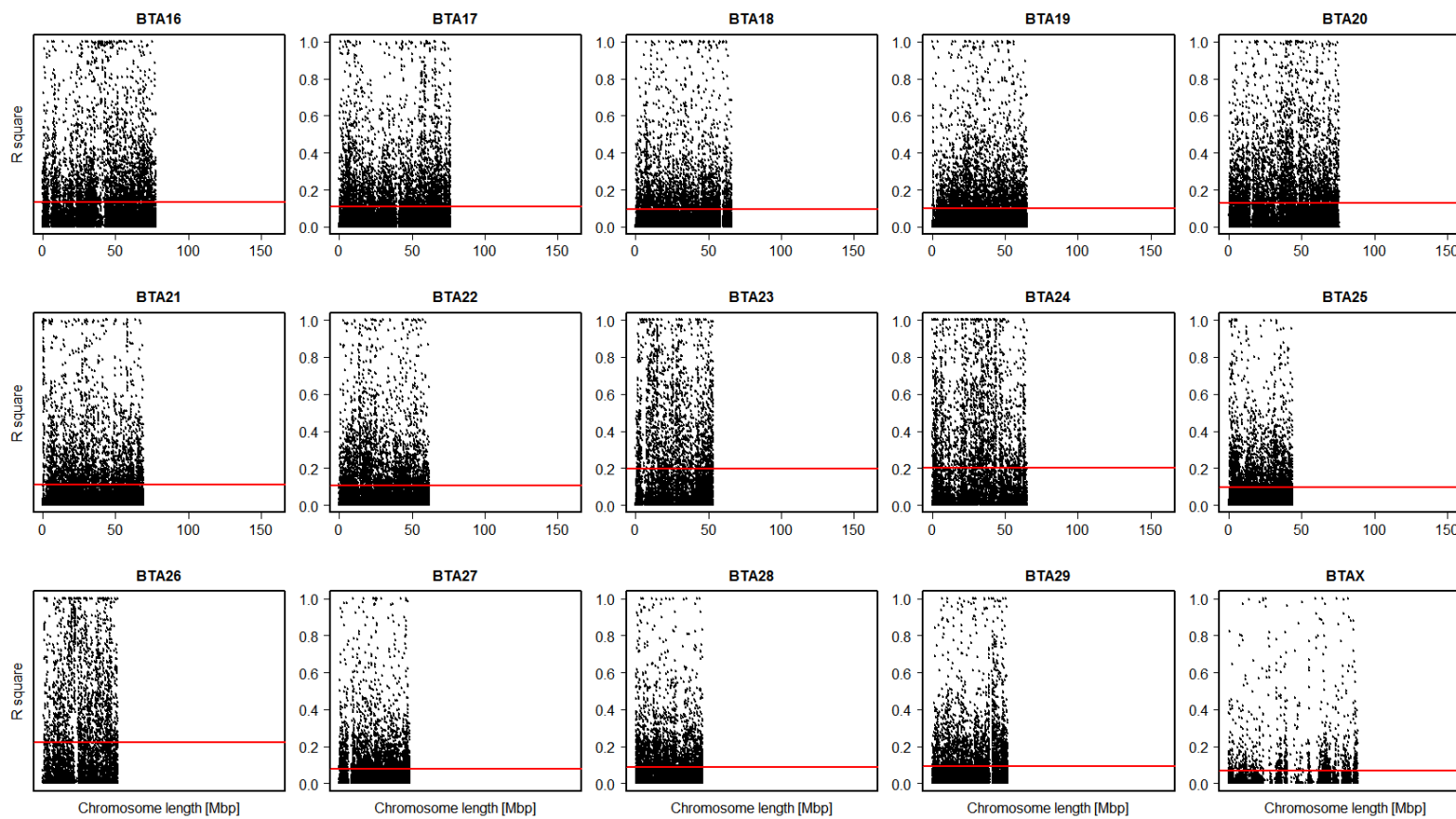


Figure 3.2: Linkage disequilibrium across chromosomes; continuation. Red line is  $r^2$  mean for each chromosome.

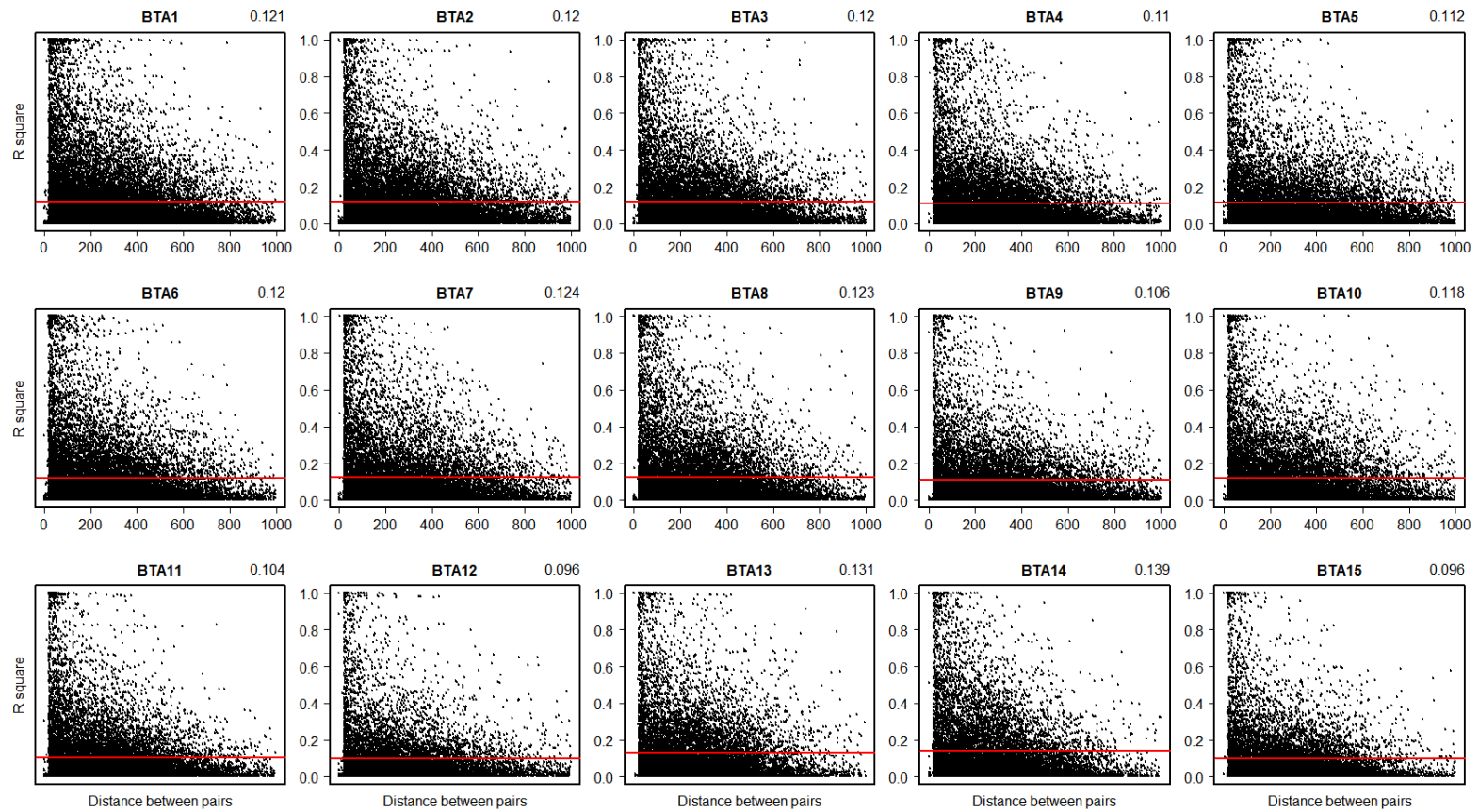


Figure 3.3: Linkage disequilibrium against distance between SNP pairs within 1 Mbp. Red line and the number in the top-right corner represents the average  $r^2$  for each chromosome.

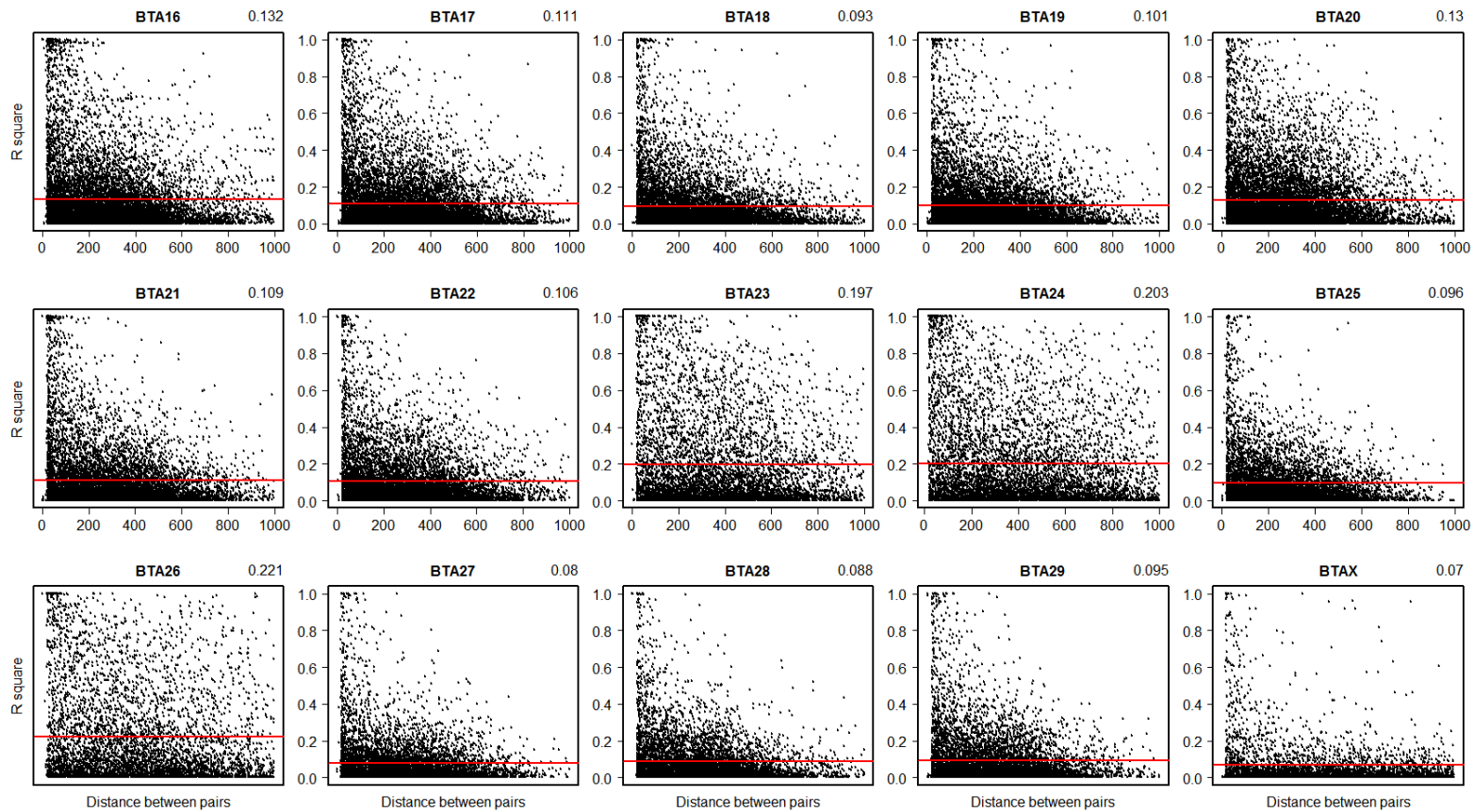


Figure 3.4: Linkage disequilibrium against distance between SNP pairs within 1 Mbp; continuation. Red line and the number in the top-right corner represents the average  $r^2$  for each chromosome.

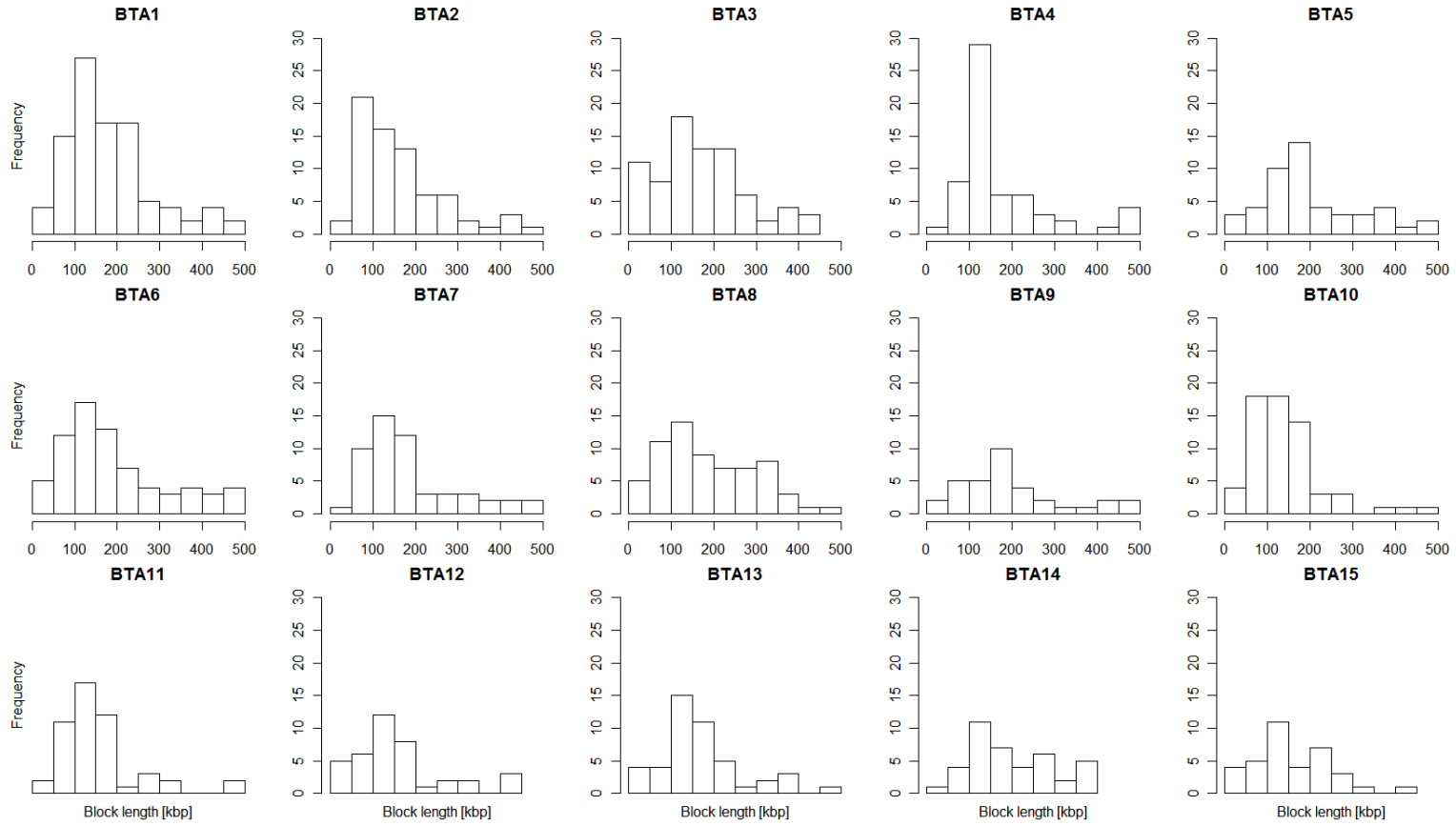


Figure 3.5: Block's length

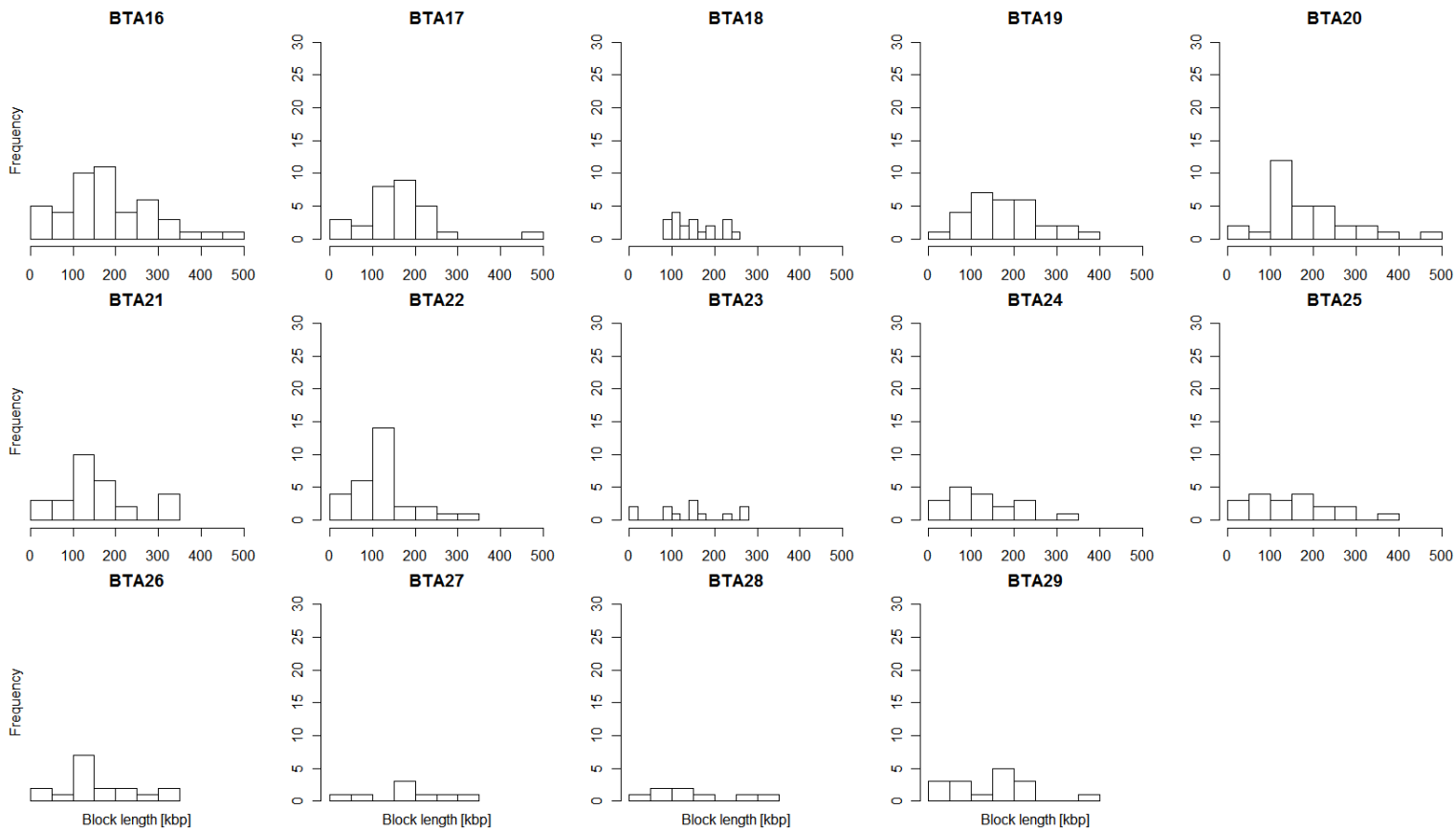


Figure 3.6: Block's length

### 3.2 TagSNP selection

The number of tagSNPs selected separately for each chromosome is presented in Table B.1 on page 57. The reduction expressing the percentage of SNPs (excluding BTAX) that are tagSNPs shows that: (i) for M1R8 it ranges from 74.5% (BTA2) to 83.2% (BTA28); (ii) for M1R5 it ranges from 60% (BTA2) to 74.6% (BTA27); (iii) for M5R8 it ranges from 67.4% (BTA16) to 77.9% (BTA28); and (iv) for M5R5 it ranges from 54.1% (BTA16) to 68.8% (BTA27). A very small number of tagSNPs are selected on BTAX. Figure 3.7 shows the reduction for all the autosomes in the four subsets. M5R5 selects the smallest subset of tagSNPs and M1R8 the largest.

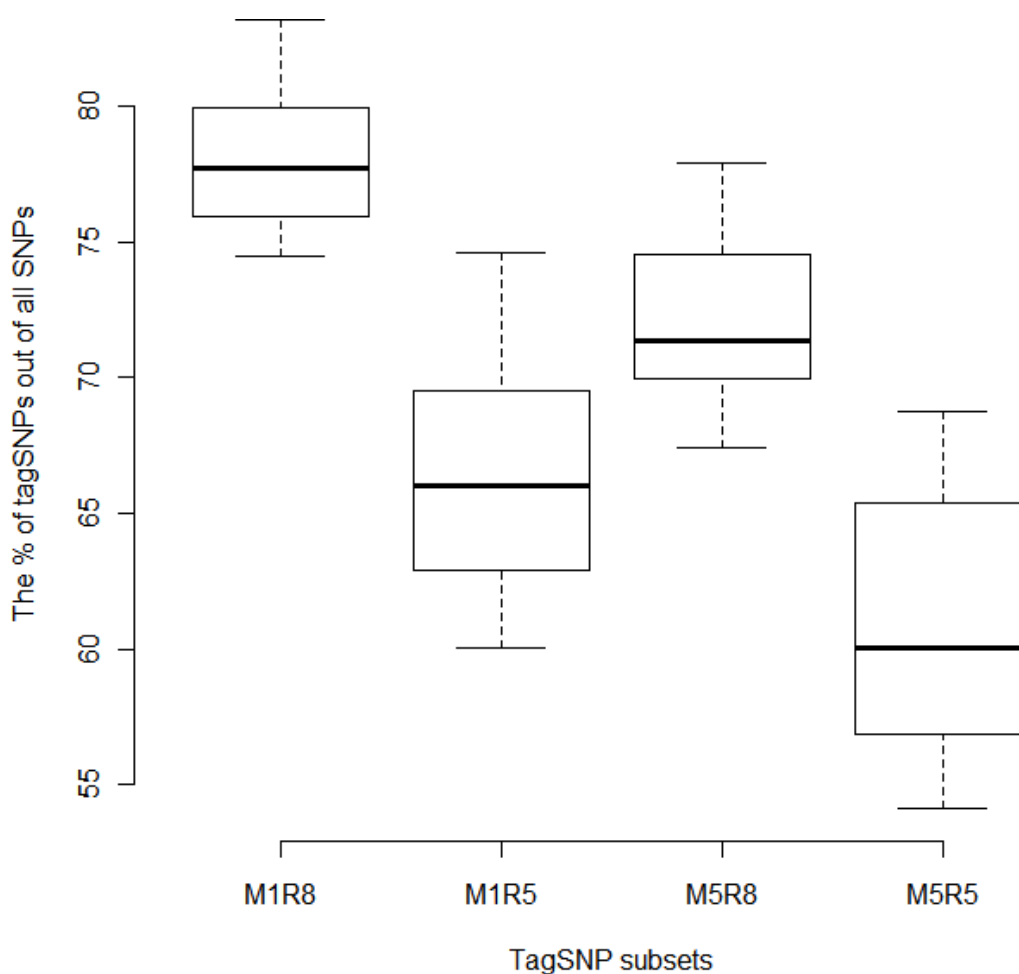


Figure 3.7: Reduction in tagSNP subsets.

TagSNP selection was also performed in six different sub-populations (see Section 2.5.1 on page 26). M1R8 was used to select the tagSNP set for each sub-population. For autosomes, the reduction varies between: (i) 74.8% (BTA7) and

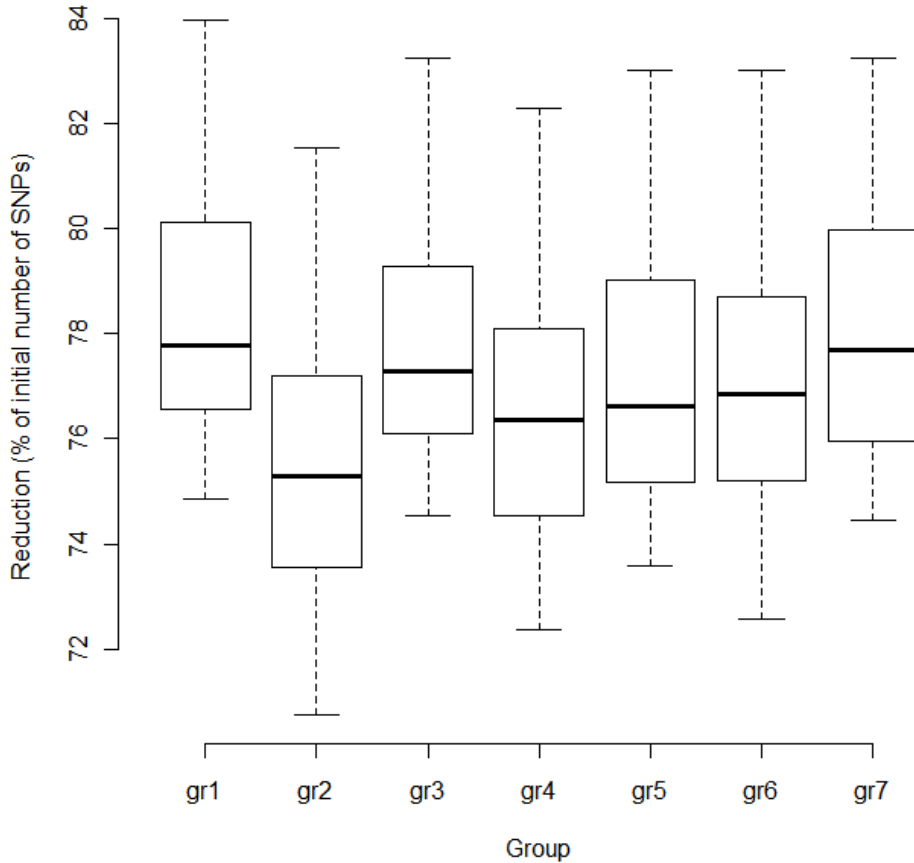


Figure 3.8: Reduction in different groups. TagSNP percentage from the original SNP set (y axis) for each group (x axis). BTAX was excluded.

84% (BTA28) for gr1; (ii) 70.7% (BTA7) and 81.5% (BTA28) for gr2; (iii) 74.5% (BTA7) and 83.2% (BTA28) for gr3; (iv) 72.4% (BTA7) and 82.3% (BTA28) for gr4; (v) 73.6% (BTA16) and 83% (BTA28) for gr5; and (vi) 72.6% (BTA7) and 83% (BTA28) for gr6. For autosomes, the overall reduction varies from 70.7% for BTA7 in gr2 to 84% for BTA28 in gr1. The average reduction is  $77.2\% \pm 2.66$ . Figure 3.8 presents the differences in reduction between sub-populations, observed for all autosomes. The largest reduction is in gr2 (38 630 tagSNPs) and the smallest overall reduction is observed for gr1 (40 267 SNPs), which is very similar for the reduction in the whole dataset (gr7). On the other hand, a distinctly larger reduction is observed for gr2 as compared to the whole dataset. Very small differences in the number of selected SNPs are observed for gr4, gr5 and gr6.

In figure 3.9, the reduction for each autosome is compared over all sub-

populations and the reference population (gr7). The regression line is plotted and the regression coefficient is shown for each comparison. The coefficient of linear regression fitted to the data varies between 0.8 (gr1-2) and 1.15 (gr2-3), and is quite close to 1, which is the expected value in the case of the reduction in each group being similar to the reduction observed for the whole data. No marked differences regarding chromosome number are observed.

In order to investigate whether the same tagSNPs are chosen in different groups, a comparison between them was performed (Figure 3.10). For autosomes, the percentage of identical tagSNPs chosen between sub-populations varies from 88.2% between gr1 and gr2 for BTA8, to 97.86% between gr3 and gr5 for BTA25. The average similarity is  $93.88\% \pm 1.69$ . The lowest similarity is observed between gr1 and gr2 (average 91.82%) and the highest is between the two random groups gr5 and gr6 (average 94.9%).

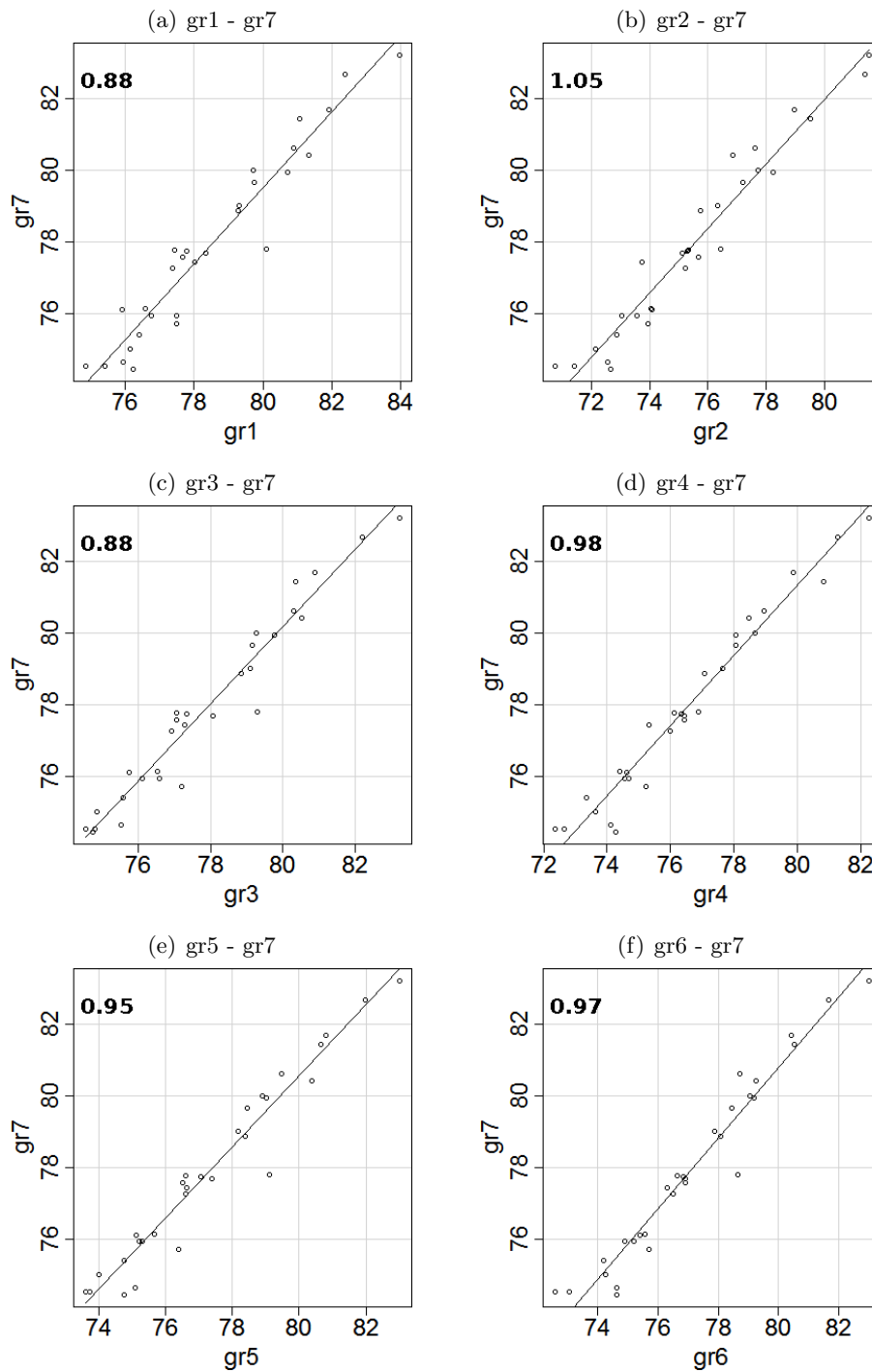


Figure 3.9: Percent of reduction in sub-population compared to the reference population. On each figure fitted linear regression line and regression coefficient are plotted.

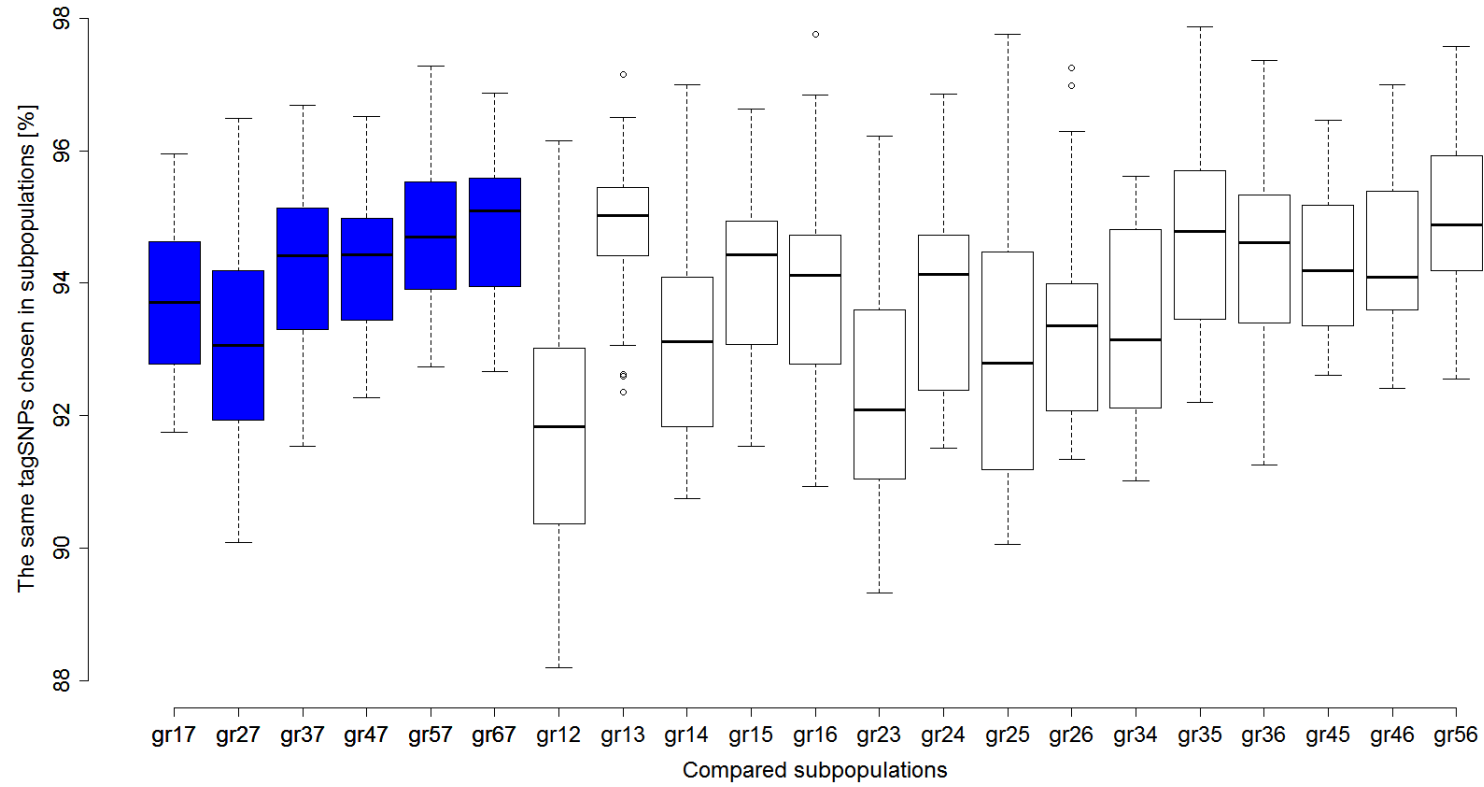


Figure 3.10: Percent of the same tagSNP between sub-populations. Only autosomes are considered. Comparisons between sub-populations and the reference population (gr7) are marked by blue.

The most important feature of a properly selected tagSNPs subset is a low LD between adjacent tagSNPs. This was verified for the largest autosome (BTA1). Figure 3.11 a) shows  $r^2$  before tagging and up to 10 neighbouring SNPs within 1 Mbp apart. There are 460 SNP pairs in high LD ( $r^2 > 0.8$ ). After tagging, high LD  $r^2$  pairs disappear from the dataset (3.11 b)). Only seven high LD pairs can be observed. A similar situation can be observed when only adjacent SNPs are considered. Figure 3.11 c) shows  $r^2$  for adjacent SNPs before tagging within 1 Mbp. After tagging, there is only one tagSNP of the 236 of high LD pairs left in the tagSNP dataset (Figure 3.11 d).

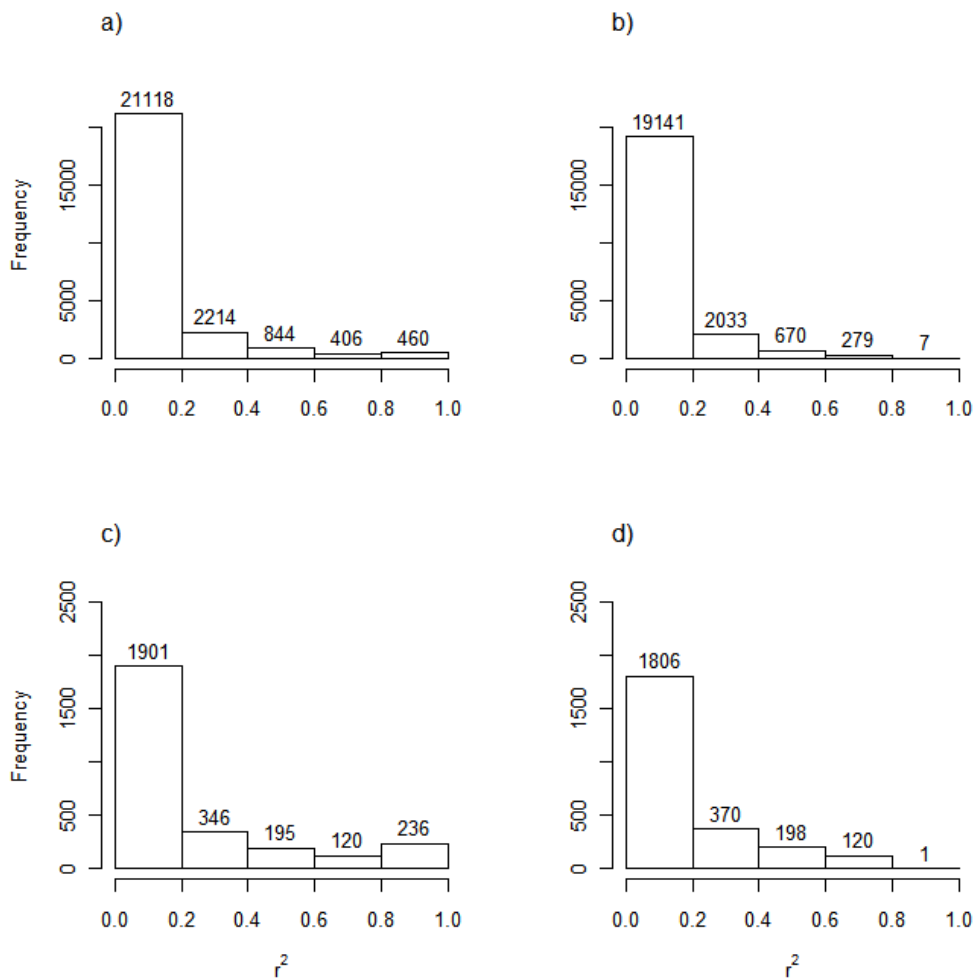


Figure 3.11: Tagging validation. Figure a) shows all pairs before tagging, lying no more than 10 SNPs from each other, while Figure b) shows tagSNPs selected from the SNPs presented in Figure a). Figure c) only shows adjacent SNPs before tagging. Figure d) shows tagSNPs selected from the SNPs presented in Figure c).

---

## Discussion

The Illumina BovineSNP50 was designed to provide SNPs which are relatively evenly distributed across the genome. *Banos and Coffey (2010)* claimed the BovineSNP50 has marginal acceptable SNP density. Moreover *Michelizzi et al. (2010)* did not observe any correlation between genes and SNPs, which is also confirmed in this study. The mean  $r^2$  calculated between maximum 10 neighbouring SNPs, located within 1 Mbp window, is 0.12. This is slightly lower (0.125) than the value reported by *Banos and Coffey (2010)*. This discrepancy in results may be caused by the different structures of the analysed populations. *Banos and Coffey (2010)* used two preselected lines of animals. In this study, a representative population is considered.

Three chromosomes (BTA23, BTA24 and BTA26) have a much higher LD than the other chromosomes and, unlike them, do not exhibit significant LD decay at short distances. These differences in LD decay for bovine chromosomes have not previously been reported so far. Only *Bohmanova et al. (2010)* mentioned higher LD on BTA7 and BTA14. The number of blocks identified in this study is much higher than the number of LD blocks reported in previous studies. In this study, 1 163 blocks were identified. This is considerably more than the 727 LD blocks identified by *Khatkar et al. (2007)* who used a smaller array, and 712 blocks identified by *Qanbari et al. (2010)* based on the same microarray used in this study. However, the mean and median block sizes obtained in this study (170 kbp, 148 kbp) are very similar to the values reported by *Qanbari et al.* (164.0 kbp, 144.0 kbp), whereas *Khatkar et al.* obtained much lower values (69.7 kbp, 3.9 kbp). Consequently, the German HF population analysed by *Qanbari et al.* has a lower genome coverage by blocks (4.67%) as compared to the Polish HF population (7.17%). Also, the Australian HF population analysed by *Khatkar et al.* had a smaller coverage (2.27%). LD dynamics can be considered long-term across tens of generations as well as short-term over a decade. The differences between the Polish and German populations may be a result of these two possible schemes. To look at these respectively: (i) with the exception of the last few

decades, these two populations have been effectively isolated (long-term scheme); and (ii) there have been different selection approaches over the last decade - more traits were selected in the German HF (short-term scheme). In the last three decades, selection has been significantly accelerated due to artificial insemination (a top bull may now have thousands of daughters in many countries).

The goal is to have the highest possible genome coverage with LD blocks. This is a measure of the quality of a microarray (*Matukumalli et al., 2009; Villa-Angulo et al., 2009*). The better the coverage, the more genetic variation captured by the blocks. Obviously, 100% coverage cannot be achieved because of factors such as crossing-over. Still, the coverage of the cattle genome is marginal. In the perfect situation, we can be certain that all genetic variation has been described when almost all the SNPs in the population have been identified. However, there is no need to put all the available SNPs into the array. Using the information provided by the linkage disequilibrium, tagSNPs can be selected from the available SNPs without substantial loss of information.

Because of the low coverage by the blocks, only pairwise tagging was used in this study. Although *Qanbari et al. (2010)* used the same array and breed, they obtained a different number of tagSNPs than was obtained in this study. It is unclear whether this was caused by the differences in the genetic structure of the populations studied, or in the methodology of tagSNP selection. *Ke et al. (2005)* distinguished between haplotype tagging and pairwise tagging as two different approaches. The approach adopted by *Qanbari et al. (2010)* was to first estimate haplotypes and then to use tagger to identify tagSNPs. This is a mixture of a haplotype and pairwise tagging. This study based its analysis entirely on the pairwise tagging approach. That is why the results of the two studies cannot be directly compared. *Khatkar et al. (2007)* and *Marques et al. (2008)* also used pairwise tagging, but smaller datasets.

It is worth mentioning that the sample size is an important factor for accurate  $r^2$  estimation and consequently for tagSNP selection. *Khatkar et al. (2008)* proposed a sample size of 50 as the smallest usable number for  $r^2$  estimation. The sample size used in this study was more than 24 times larger. Reliable tagSNP selection was therefore ensured in this study.

This is the first time that population parameters, such as the degree of inbreeding and relationship, have been factored into tagging analysis. This study revealed that the vast majority of tagSNPs were shared between different subpopulations (gr1-gr6).

Tagging is a crucial part of SNP assay (*Ke et al., 2005*). It results in significant time and cost savings. Smaller arrays are cheaper to construct and easier to work with (*Ke et al., 2005*). The downside of tagging is the possibility that some variation may not be covered by tagSNPs. Since the first genome-wide linkage disequilibrium analysis of the bovine genome based on SNPs was performed by

*Khatkar et al.* (2007), it has been clear that the number of SNPs available for cattle is not sufficient to cover the entire variance in the cattle genome (e.g. single nucleotide variance, copy number variance, microsatellite variance). *Khatkar et al.* proposed 30 - 50 kbp spacing between 2 tagSNPs to achieve full variation coverage. They assumed that a microarray consisting of 200 000 to 250 000 SNPs would be sufficient. Those authors worked with an Affymetrix 10K array with additional SNPs. No common agreement has been reached regarding the ideal number of SNPs required for both GWAS and LD block coverage. Although *Gautier et al.* (2007) worked on a set of SNPs that was ten times smaller, they proposed a microarray containing 1 tagSNP per 10 kbp as this would result in approximately 300 000 tagSNPs. This observation was later supported by *Kim and Kirkpatrick* (2009) and now in this study. *McKay et al.* (2007), *Prasad et al.* (2008) and *Sargolzaei et al.* (2008) arrived at completely different conclusions. They claimed that the minimum array for association study should consist of SNPs less than 70 kbp apart. This would result in approximately 50 000 SNP. *Villa-Angulo et al.* (2009), working on a much larger set of SNPs, proposed an assay of 30 000 SNPs for GWAS and 580 000 to construct the whole-genome haplotype-block structure. *Qanbari et al.* (2010) suggested that 75 000 SNPs with  $MAF \geq 0.05$  or 50 000 SNPs with  $MAF \geq 0.15$  is useful for GWAS. For genomic evaluation, *Banos and Coffey* (2010) and *de Roos et al.* (2008) proposed 300 000 SNPs. However, this number is dependent on trait heritability and mode (pure polygenic, oligogenic). It is very difficult, however, to estimate the number of SNPs required for a near perfect genome-wide variance coverage. In concurrence with the authors cited above and the results of this study, the following simple rule may prove helpful: the more SNPs that can be discovered in the population, the fewer the tagSNPs that can be allocated to a near perfect array. To recapitulate, the larger the number of available SNPs, the more accurate the haplotype block definition through LD estimation. This is illustrated in Figure 4.1 which shows the LD blocks for a 3 Mbp fragment of the cattle genome and a 0.5 Mbp fragment of the human genome. Although the human fragment is 6 times shorter, it contains many more SNPs than the cattle fragment. LD blocks can therefore be more precisely identified. Furthermore, the LD pattern is clear and easy to interpret, and the coverage by blocks is greater. Nowadays, two dense bovine SNPs panels are commercially available. Affymetrix has an array of 640 000 SNPs preselected from 3 000 000 SNPs. Illumina has a larger array consisting of 777 000 SNPs selected from polymorphisms identified for more than 20 breeds.

Linkage disequilibrium can be affected not only by crossing-over, but also by selection, mutation, genetic drift, and migration (*Hardy, 1908; Weinberg, 1908*). The widespread availability of genomic data, in combination with steadily increasing computational power, allows for a precise *in silico* description of genome

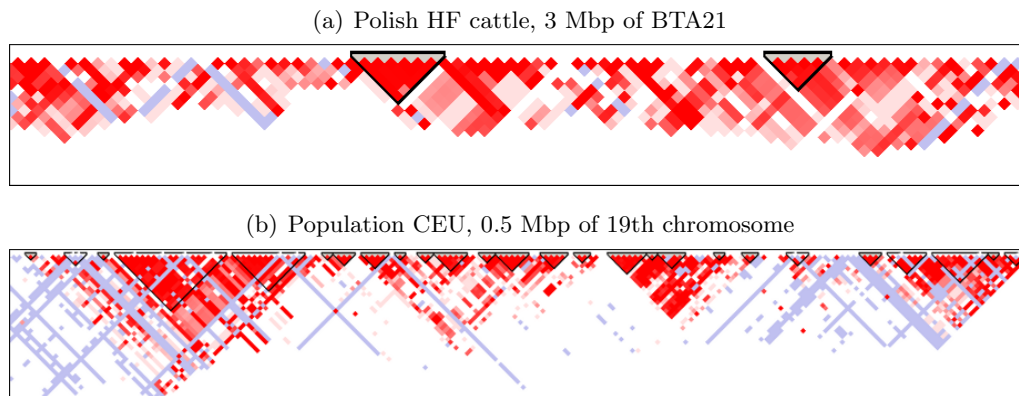


Figure 4.1: LD in human and cattle populations quantified by  $D'$ . Dark red and dark grey represent  $D' = 1$ . Pink, bright red and white represent  $D' < 1$ . LD blocks are marked by a black border.

dynamics, such as LD structure, which was considered in this study. LD analysis allows for monitoring of historical recombination. By comparing different subsets of data, this analysis shows that the pairwise tagSNP selection approach based on LD information is an universal procedure which is independent of the reference population structure.

---

# Bibliography

- Affymetrix Inc. “Affymetrix megallele genechip bovine 10k snp array.” (2005)  
[cited at p. 7, 12]
- Allen A.R., Taylor M., McKeown B., Curry A.I., Lavery J.F., Mitchell A., Hartshorne D., Fries R., Skuce R.A. “Compilation of a panel of informative single nucleotide polymorphisms for bovine identification in the northern irish cattle population.” *BMC Genetics*, 11 (2010) [cited at p. 6]
- Banos G., Coffey M.P. “Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows.” *Journal of Dairy Science*, 93(6):2775–2778 (2010) [cited at p. 12, 19, 43, 45]
- Barendse W., Reverter A., Bunch R.J., Harrison B.E., Barris W., Thomas M.B. “A validated whole-genome association study of efficient food conversion in cattle.” *Genetics*, 176:1893–1905 (2007) [cited at p. 6]
- Barrett J.C. “Haploview: Visualization and analysis of snp genotype data.” *CSH Protoc.*, 10:1101 (2009) [cited at p. 25]
- Barrett J.C., Fry B., Maller J. “Haploview: analysis and visualization of ld and haplotype maps.” *Bioinformatics*, 21(2):263–265 (2004) [cited at p. 10, 24, 25]
- Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Rapp B.A., Wheeler D.L. “Genbank.” *Nucleic Acid Research*, 28:15–18 (2000) [cited at p. 5]
- Boehnke M. “A look at linkage disequilibrium.” *Nature*, 25:246–247 (2000)  
[cited at p. 8]
- Bohmanova J., Sargolzaei M., Schenkel F.S. “Characteristics of linkage disequilibrium in north american holsteins.” *BMC Genomics*, 11 (2010) [cited at p. 12, 19, 43]
- Brookes A.J. “The Essence of SNPs.” *Gene*, 234:177–186 (1999) [cited at p. 5, 6, 7]

- Carlson C.S., Eberle M.A., Rieder M.J. “Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.” *Am. J. Hum. Genet.*, 74:106—120 (2004) [cited at p. 9, 24, 26]
- de Bakker P.I. “Selection and evaluation of tag-snps using tagger and hapmap.” *CSH Protoc.*, 10:1101 (2009) [cited at p. 25]
- de Bakker P.I., Yelensky R., Pe’er I., Gabriel S.B. “Efficiency and power in genetic association studies.” *Nat. Genet.*, 37(11):1217–1223 (2005) [cited at p. 25]
- de Roos A.P.W., Hayes B.J., Spelman R.J., Goddard M.E. “Linkage disequilibrium and persistence of phase in holstein–friesian, jersey and angus cattle.” *Genetics*, 179:1503—1512 (2008) [cited at p. 12, 45]
- Devlin B., Risch N. “A comparison of linkage disequilibrium measures for fine-scale mapping.” *Genomics*, 29:311–322 (1995) [cited at p. 24]
- Ding K., Kullo I.J. “Methods for the selection of tagging snps: a comparison of tagging efficiency and performance.” *European Journal of Human Genetics*, 15:228—236 (2007) [cited at p. 9]
- Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander R.S., Daly M.J., Altshuler D. “The structure of haplotype blocks in the human genome.” *Science*, 296:2225–2229 (2002) [cited at p. 24]
- Gautier M., Faraut T., Moazami-Goudarzi K., Navratil V., Foglio M., Grohs C., Boland A., Garnier J., Boichard D., Lathrop G.M., Gut I.G., Eggen A. “Genetic and haplotypic structure in 14 european and african cattle breeds.” *Genetics*, 177:1059—1070 (2007) [cited at p. 13, 45]
- Goldstein D.B., Weale M.E. “Population genomics: Linkage disequilibrium holds the key.” *Current Biology*, 11:R576—R579 (2001) [cited at p. 8]
- Hardenbol P., Yu F.L., Belmont J., MacKenzie J., Bruckner C. “Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted snps genotyped in a single tube assay.” *Genome Research*, 15:269—275 (2005) [cited at p. 12]
- Hardy G.H. “Mendelian proportions in a mixed population.” *Science*, 28(706):49–50 (1908) [cited at p. 45]
- Hill W.G., Weir B.S. “Maximum-likelihood estimation of gene location by linkage disequilibrium.” *Am. J. Hum. Genet.*, 54:705–714 (1994) [cited at p. 24, 55]
- Illumina Inc. “Illumina bovinesnp50 54’001 array.” (2009) [cited at p. 12, 15]

- International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome." *Nature*, 409:860—921 (2001) [cited at p. 5]
- Johnson G.C., Esposito L., Barratt B.J. "Haplotype tagging for the identification of common disease genes." *Nature Genetics*, 29:233—237 (2001) [cited at p. 9]
- Ke X., Miretti M.M., Broxholme J., Hunt S., Beck S., Bentley D.R., Deloukas P., Cardon L.R. "A comparison of tagging methods and their tagging space." *Human Molecular Genetics*, 14:2757—2767 (2005) [cited at p. 44]
- Khatkar M.S., Nicholas F.W., Collins A.R., Zenger K.R., Cavanagh J.A.L., Barris W., Schnabe R.D., Taylor J.F., Raadsma H.W. "Extent of genome-wide linkage disequilibrium in australian holstein-friesian cattle based on a high-density snp panel." *BMC Genomics*, 9 (2008) [cited at p. 12, 44]
- Khatkar M.S., Zenger K.R., Hobbs M., Hawken R.J., Cavanagh J.A.L., Barris W., McClintock A.E., McClintock S., Thomson P.C., Tier B., Nicholas F.W., Raadsma H.W. "A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in holstein-friesian cattle." *Genetics*, 176:763—772 (2007) [cited at p. 12, 13, 43, 44, 45]
- Kim E.S., Kirkpatrick B.W. "Linkage disequilibrium in the north american holstein population." *Animal Genetics*, 40:279—288 (2009) [cited at p. 8, 12, 45]
- LaFramboise T. "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances." *Nucleic Acids Research*, 37:4181—4193 (2009) [cited at p. 6, 7, 11]
- Lewin H.A. "It's a bull's market." *Science*, 324:478—479 (2009) [cited at p. 11]
- Lewontin R.C., Kojima K. "The evolutionary dynamics of complex polymorphisms." *Evolution*, 14(4):458—472 (1960) [cited at p. 24]
- Mah J.T., Chia K.S. "A gentle introduction to snp analysis: resources and tools." *Journal of Bioinformatics and Computational Biology*, 5:1123—38 (2007) [cited at p. 6]
- Marques E., Schnabel R.D., Stothard P., Kolbehdari D., Wang Z., Taylor J.F., Moore S.S. "High density linkage disequilibrium maps of chromosome 14 in holstein and angus cattle." *BMC Genetics*, 9 (2008) [cited at p. 13, 44]
- Matukumalli L.K., Lawley C.T., Schnabel R.D., Taylor J.F., Allan M.F., Heaton M.P., O'Connell J., Moore S.S., Smith T.P.L., Sonstegard T.S., Van Tassell C.P. "Development and characterization of a high density snp genotyping assay for cattle." *PLoS ONE*, 4(4):e5350 (2009) [cited at p. 7, 8, 11, 12, 15, 19, 44]

- McCarthy M.I., Abecasis G.R., Cardon L.R., Goldstein D.B., Little J., Ioannidis J.P.A., Hirschhorn J.N. “Genome-wide association studies for complex traits: consensus, uncertainty and challenges.” *Nature Reviews Genetics*, 9:356–369 (2008) [cited at p. 5]
- McKay S.D., Schnabe R.D., Murdoch B.M., Matukumalli L.K., Aerts J., Coppeters W., Crews D., Neto E.D., Gill C.A., Gao C., Mannen H., Stothard P., Wang Z., Van Tassell C.P., Williams J.L., F T.J., Moore S.S. “Whole genome linkage disequilibrium maps in cattle.” *BMC Genetics*, 8 (2007) [cited at p. 12, 45]
- Michelizzi V.N., Wu X., Dodson M.V., J M.J., Zambrano-Varon J., McLean D.J., Jiang Z. “A global view of 54,001 single nucleotide polymorphisms (snps) on the illumina bovinesnp50 beadchip and their transferability to water buffalo.” *International Journal of Biological Sciences*, 7(1):18–27 (2010) [cited at p. 43]
- Oliphant A., Barker D.L., Stuelpnagel J.R., Chee M.S. “Beadarray technology: enabling an accurate, cost-effective approach to high-throughput genotyping.” *BioTechniques*, 32:S56–S61 (2002) [cited at p. 7]
- Prasad A., Schnabel R.D., McKay S.D., Murdoch B., Stothard P., Kolbehdari D., Wang Z., Taylor J.F., Moore S.S. “Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle.” *Animal Genetics*, 39:597–605 (2008) [cited at p. 13, 45]
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P. “Plink: a toolset for whole-genome association and population-based linkage analysis.” (2007). [Http://pngu.mgh.harvard.edu/purcell/plink/](http://pngu.mgh.harvard.edu/purcell/plink/) [cited at p. 24]
- Qanbari S., Pimentel E.C.G., Tetens† J., Thaller G., Lichtner P., Shari A.R., Simianer H. “The pattern of linkage disequilibrium in german holstein cattle.” *Animal Genetics*, 41:346–56 (2010) [cited at p. 12, 13, 19, 43, 44, 45]
- Sargolzaei M., Schenkel F.S., Jansen G.B., Schaeffer L.R. “Extent of linkage disequilibrium in holstein cattle in north america.” *Journal of Dairy Science*, 91:2106—2117 (2008) [cited at p. 12, 45]
- Schaeffer L.R. “Strategy for applying genome-wide selection in dairy cattle.” *Journal of Animal Breeding and Genetics*, 123:218–223 (2006) [cited at p. 6]
- Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. “dbSNP: the NCBI database of genetic variation.” *Nucleic Acid Research*, 29:308–311 (2001) [cited at p. 5]

- The 1000 Genomes Project Consortium. “A map of human genome variation from population-scale sequencing.” *Nature*, 467:1061—1073 (2010) [cited at p. 5]
- The Bovine Genome Sequencing and Analysis Consortium, Elvik C.G., Tellam R.L., Worley K.C. “The genome sequence of taurine cattle: A window to ruminant biology and evolution.” *Science*, 324:522–528 (2009) [cited at p. 5, 11, 13, 15, 18]
- The Bovine HapMap Consortium. “Genome-wide survey of snp variation uncovers the genetic structure of cattle breeds.” *Science*, 324:528–532 (2009) [cited at p. 7, 11]
- The International HapMap Consortium. “The international hapmap project.” *Nature*, 426:789–796 (2003) [cited at p. 5, 7]
- Thorisson G.A., Smith A.V., Krishnan L., Stein L.D. “The international hapmap project web site.” *Genome Research*, 15:1592—1593 (2005) [cited at p. 7, 10]
- Venter J.C., et al. “The sequence of the human genome.” *Science*, 291:1304—1351 (2001) [cited at p. 5]
- Villa-Angulo R., Matukumalli L.K., Gill C.A., Choi J., Van Tassell C.P., Grefenstette J.J. “High-resolution haplotype block structure in the cattle genome.” *BMC Genetics*, 10 (2009) [cited at p. 13, 44, 45]
- Weinberg W. “Über den nachweis der vererbung beim menschen.” *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 64:368—382 (1908) [cited at p. 45]
- Zimin A.V., Delcher A.L., Florea L. “A whole-genome assembly of the domestic cow, *bos taurus*.” *Genome Biol.*, (10)4:R42 (2009) [cited at p. 15]



# Appendices



---

## LD measures

Following *Hill and Weir (1994)* two-allele model was used to explain the measures of LD used in analysis below. For each allele (1 and 2), and for *loci A* and *B* let *AB* denote as two-locus haplotype. Then each haplotype frequency can be denoted as follow:

Haplotype	Frequency
$A_1B_1$	$x_{11}$
$A_1B_2$	$x_{12}$
$A_2B_1$	$x_{21}$
$A_2B_2$	$x_{22}$

Then allele frequencies can be described as follow:

Allele	Frequency
$A_1$	$p1 = x_{11} + x_{12}$
$A_2$	$p2 = x_{21} + x_{22}$
$B_1$	$q1 = x_{21} + x_{11}$
$B_2$	$q2 = x_{22} + x_{12}$

When these haplotypes are in linkage equilibrium, than the haplotype frequencies can be estimated as follow:

$$\begin{aligned}x_{11} &= p_1q_1 \\x_{21} &= p_2q_1 \\x_{12} &= p_1q_2 \\x_{22} &= p_2q_2\end{aligned}$$

However, if linkage disequilibrium occurs, these frequencies need to be reevaluated as follow:

$$\begin{aligned}x_{11} &= p_1q_1 + D \\x_{21} &= p_2q_1 - D\end{aligned}$$

$$\begin{aligned}x_{12} &= p_1q_2 - D \\x_{22} &= p_2q_2 + D\end{aligned}$$

In this case,  $D$  is the exception of the Hardy-Weinberg equilibrium. Unfortunately, the value of  $D$  depends on the frequency of alleles. Therefore  $D$  can not be observed if the frequency of one of the alleles is equal to 1 or 0. Moreover, if allele frequencies are equal to 0.5,  $D$  takes the maximum value.

---

## TagSNP selection

Table B.1: TagSNP subsets.

Chromosome	SNPs	M1R8	%	M1R5	%	M5R8	%	M5R5	%
BTA1	3343	2496	74.7	2039	61.0	2303	68.9	1862	55.7
BTA2	2764	2058	74.5	1710	61.9	1899	68.7	1559	56.4
BTA3	2566	1987	77.4	1660	64.7	1803	70.3	1491	58.1
BTA4	2541	1930	76.0	1652	65.0	1787	70.3	1525	60.0
BTA5	2181	1661	76.2	1414	64.8	1545	70.8	1304	59.8
BTA6	2535	1920	75.7	1592	62.8	1771	69.9	1452	57.3
BTA7	2294	1710	74.5	1400	61.0	1558	67.9	1261	55.0
BTA8	2362	1794	76.0	1486	62.9	1628	68.9	1332	56.4
BTA9	2036	1606	78.9	1391	68.3	1465	72.0	1259	61.8
BTA10	2179	1693	77.7	1426	65.4	1561	71.6	1302	59.8
BTA11	2267	1763	77.8	1524	67.2	1656	73.0	1426	62.9
BTA12	1683	1309	77.8	1136	67.5	1212	72.0	1045	62.1
BTA13	1802	1359	75.4	1113	61.8	1261	70.0	1018	56.5
BTA14	1722	1292	75.0	1034	60.0	1187	68.9	935	54.3
BTA15	1688	1345	79.7	1191	70.6	1258	74.5	1104	65.4
BTA16	1606	1197	74.5	971	60.5	1083	67.4	869	54.1
BTA17	1585	1275	80.4	1078	68.0	1165	73.5	973	61.4
BTA18	1351	1081	80.0	952	70.5	1014	75.1	885	65.5
BTA19	1378	1089	79.0	958	69.5	1030	74.7	901	65.4
BTA20	1564	1261	80.6	1007	64.4	1132	72.4	889	56.8
BTA21	1419	1080	76.1	921	64.9	1003	70.7	849	59.8
BTA22	1299	1004	77.3	861	66.3	926	71.3	786	60.5
BTA23	1083	885	81.7	776	71.7	825	76.2	716	66.1
BTA24	1294	1004	77.6	854	66.0	920	71.1	774	59.8
BTA25	987	804	81.5	700	70.9	759	76.9	656	66.5
BTA26	1086	845	77.8	719	66.2	775	71.4	652	60.0
BTA27	977	808	82.7	729	74.6	749	76.7	672	68.8
BTA28	942	784	83.2	691	73.4	734	77.9	643	68.3
BTA29	1048	838	80.0	746	71.2	785	74.9	695	66.3
BTAX	747	47	6.3	45	6.0	43	5.8	41	5.5

Table B.2: The reduction in each subpopulation

Chromosome	gr1	gr2	gr3	gr4	gr5	gr6	gr7
BTA1	75.9	72.5	75.5	74.1	75.1	74.6	74.7
BTA2	76.2	72.6	74.7	74.3	74.7	74.6	74.5
BTA3	78.0	73.7	77.3	75.3	76.6	76.3	77.4
BTA4	76.7	73.6	76.1	74.7	75.3	75.2	76.0
BTA5	76.6	74.0	76.5	74.4	75.7	75.6	76.2
BTA6	77.5	73.9	77.2	75.2	76.4	75.7	75.7
BTA7	74.8	70.7	74.5	72.4	73.7	72.6	74.5
BTA8	77.5	73.0	76.6	74.6	75.2	74.9	76.0
BTA9	79.3	75.7	78.8	77.1	78.4	78.1	78.9
BTA10	78.3	75.1	78.1	76.5	77.4	76.9	77.7
BTA11	77.8	75.3	77.3	76.4	77.1	76.8	77.8
BTA12	77.4	75.3	77.1	76.1	76.6	76.6	77.8
BTA13	76.4	72.9	75.6	73.4	74.8	74.2	75.4
BTA14	76.1	72.1	74.9	73.6	74.0	74.3	75.0
BTA15	79.7	77.2	79.1	78.1	78.4	78.4	79.7
BTA16	75.4	71.4	74.8	72.7	73.6	73.0	74.5
BTA17	81.3	76.8	80.5	78.5	80.4	79.2	80.4
BTA18	79.7	77.7	79.3	78.7	78.9	79.1	80.0
BTA19	79.3	76.3	79.1	77.6	78.2	77.9	79.0
BTA20	80.9	77.6	80.3	79.0	79.5	78.7	80.6
BTA21	75.9	74.1	75.8	74.6	75.1	75.4	76.1
BTA22	77.4	75.2	76.9	76.0	76.6	76.5	77.3
BTA23	81.9	78.9	80.9	79.9	80.8	80.4	81.7
BTA24	77.7	75.7	77.0	76.4	76.5	76.9	77.6
BTA25	81.1	79.5	80.3	80.9	80.6	80.5	81.5
BTA26	80.1	76.4	79.3	76.9	79.1	78.6	77.8
BTA27	82.4	81.4	82.2	81.3	82.0	81.7	82.7
BTA28	84.0	81.5	83.2	82.3	83.0	83.0	83.2
BTA29	80.7	78.2	79.8	78.1	79.0	79.2	80.0
BTAX	6.3	6.2	6.3	6.2	6.3	6.2	6.3

Table B.3: The percent of identical tagSNPs chosen within the seven groups of animals.

	gr12	gr13	gr14	gr15	gr16	gr17	gr23	gr24	gr25	gr26	gr27	gr34	gr35	gr36	gr37	gr45	gr46	gr47	gr56	gr57	gr67
BTA1	89.6	93.1	91.4	92.6	92.1	92.2	89.6	91.8	91.1	91.3	91.1	92.1	92.7	93.2	92.8	92.9	93.1	92.9	93.3	93.7	94.0
BTA2	90.4	92.4	92.4	92.7	92.8	92.8	91.0	92.7	91.2	91.6	91.9	93.2	93.9	94.9	93.7	93.7	94.4	93.4	94.2	94.4	94.5
BTA3	88.7	93.7	91.3	92.4	93.1	92.3	89.8	91.7	90.8	91.4	90.8	91.8	93.4	92.8	92.8	92.8	93.0	92.9	93.4	93.8	93.2
BTA4	91.5	94.4	92.6	93.1	92.8	93.1	91.9	93.6	92.6	93.1	92.7	92.8	93.7	93.9	93.2	94.3	94.0	93.6	94.2	94.4	94.0
BTA5	92.5	95.3	93.1	93.9	94.6	93.7	92.1	94.5	92.6	94.0	94.5	92.6	94.5	95.0	93.7	93.9	93.6	94.1	94.8	94.5	95.1
BTA6	89.5	95.2	91.5	93.6	92.0	92.7	89.6	92.1	91.0	92.1	92.0	91.6	93.5	92.5	93.3	92.8	92.8	93.3	95.3	93.3	94.0
BTA7	88.9	94.1	91.7	92.8	91.1	92.1	90.0	91.5	90.1	91.5	91.1	91.1	93.3	92.2	93.5	92.6	92.9	92.8	92.5	93.9	92.7
BTA8	88.2	92.6	91.0	91.6	90.9	91.7	89.3	91.7	90.8	91.4	91.0	91.5	92.2	92.0	92.1	92.9	95.7	92.3	93.1	93.3	92.7
BTA9	91.4	95.5	92.5	95.1	94.1	94.3	91.5	93.7	92.6	92.1	93.3	93.1	95.0	95.5	94.9	93.6	93.8	93.6	95.1	94.7	95.1
BTA10	91.2	96.1	93.2	93.7	93.8	93.6	91.2	94.0	92.0	92.6	92.8	93.1	93.7	93.5	93.5	93.4	94.3	94.3	94.2	94.3	94.4
BTA11	92.3	94.8	93.8	94.3	94.4	94.3	93.0	94.2	93.2	93.3	93.0	94.8	96.1	95.2	94.8	94.3	94.1	94.4	96.7	95.4	95.6
BTA12	93.2	95.8	94.1	95.8	94.9	95.2	93.5	94.8	93.8	94.3	93.9	94.3	96.1	95.7	95.3	95.3	96.0	95.3	95.8	95.7	95.2
BTA13	91.0	94.6	91.4	94.0	93.1	93.8	91.8	95.2	92.8	93.8	92.6	92.2	94.8	94.3	95.0	93.2	94.1	93.5	94.3	94.6	94.6
BTA14	89.2	92.6	91.8	91.5	92.2	92.1	91.1	91.6	90.4	91.7	90.9	91.4	92.5	93.4	91.5	93.7	92.4	93.4	93.6	92.7	93.4
BTA15	93.0	95.4	94.4	94.7	94.7	94.4	93.6	94.4	95.2	94.3	94.5	94.8	96.0	94.6	95.1	94.9	95.4	94.9	95.7	95.2	95.6
BTA16	89.5	93.1	90.8	92.2	91.4	92.2	89.9	91.9	91.1	91.4	90.1	91.0	93.3	91.3	92.3	92.9	92.7	92.9	92.6	93.8	93.0
BTA17	91.1	94.5	92.8	94.7	93.1	93.0	90.9	92.7	91.5	92.6	92.4	92.1	94.4	94.2	93.3	96.1	95.3	94.7	94.2	93.6	95.4
BTA18	93.9	95.5	95.0	94.8	94.8	95.4	94.4	94.7	94.5	93.5	93.7	95.6	95.4	95.0	94.7	95.3	96.3	95.0	96.0	97.3	96.2
BTA19	92.5	94.7	94.4	94.8	93.8	94.0	93.3	94.7	93.4	93.8	94.1	94.5	95.7	95.2	96.1	95.2	95.2	95.6	95.3	95.6	95.3
BTA20	90.9	94.5	92.8	93.4	92.3	93.1	91.1	92.4	91.5	93.3	91.7	93.2	92.8	92.7	93.3	93.9	93.6	93.7	93.1	94.7	93.3
BTA21	92.1	94.4	93.8	94.4	94.7	93.6	93.8	94.4	94.4	93.4	94.2	93.5	95.0	94.9	94.4	94.2	93.8	94.9	94.5	94.9	95.5
BTA22	92.8	95.3	93.9	94.9	94.2	94.6	93.6	94.0	93.8	93.8	93.1	94.4	95.7	94.2	95.2	94.8	94.8	95.0	95.6	95.5	95.5
BTA23	94.8	95.3	94.7	96.2	95.0	95.0	93.8	95.4	95.7	95.1	94.4	95.2	96.7	96.5	96.7	96.5	96.7	95.9	96.7	96.1	96.8
BTA24	92.5	95.0	94.0	95.1	94.3	94.1	93.0	94.6	94.5	93.5	94.4	93.7	94.4	96.1	94.0	94.0	94.0	94.5	94.9	95.1	93.4
BTA25	95.5	96.5	97.0	96.6	97.8	95.5	96.2	95.9	95.5	97.0	95.7	95.5	97.9	97.4	96.1	95.9	96.6	96.4	96.7	95.7	96.7
BTA26	91.8	95.7	92.4	94.9	94.5	93.6	92.7	94.1	94.1	93.8	93.6	93.0	95.6	94.3	94.7	94.3	92.6	94.4	95.9	94.1	94.6
BTA27	96.1	97.1	94.9	96.5	96.1	95.8	95.9	96.9	97.8	97.2	96.5	95.0	96.1	96.5	96.4	96.4	97.0	95.4	97.0	96.3	96.9
BTA28	95.2	96.2	95.6	96.1	96.8	96.0	94.8	95.2	95.9	96.3	95.6	95.5	97.4	96.4	96.3	95.9	96.2	96.5	97.6	96.2	96.2
BTA29	93.5	95.4	93.6	94.4	95.3	94.9	94.4	96.5	94.9	94.5	94.2	95.1	95.0	95.3	94.6	94.2	95.2	94.8	96.4	95.5	96.1
BTAX	97.9	100.0	100.0	100.0	97.9	100.0	97.9	97.9	97.9	97.8	97.9	100.0	100.0	97.9	100.0	100.0	97.9	100.0	97.9	100.0	97.9



---

## List of Symbols and Abbreviations

Abbreviation	Description
BTA	<i>Bos taurus</i> chromosome
CEU	One of the HapMap population; it can be referred as northern and western European ancestors living in Utah (USA)
GWAS	Genome-Wide Association Study
HF	Holstein-Friesian dairy cattle breed
MAF	Minor Allele Frequency
SNP	Single Nucleotide Polymorphism
YRI	One of the HapMap population from Ibadan, Nigeria, from individuals who identified themselves as having four Yoruba grandparents

---

---

## List of Figures

1.1	Illumina BovineSNP50 microarray . . . . .	8
1.2	Linkage Disequilibrium in two populations . . . . .	9
1.3	Linkage Disequilibrium in two human populations . . . . .	10
1.4	<i>Bos taurus</i> karyotype . . . . .	11
2.1	Coefficient of relationship for all bulls . . . . .	16
2.2	Average coefficient of relationship and inbreeding . . . . .	16
2.3	Scatter plots of chromosome length, the number of genes and SNPs . . . . .	18
2.4	SNP distribution . . . . .	20
2.5	SNP distribution; continuation . . . . .	21
2.6	Distances between SPNs . . . . .	22
2.7	Distances between SPNs; continuation . . . . .	23
2.8	Genome-wide Distribution of MAF . . . . .	25
2.9	MAF on each chromosome . . . . .	26
3.1	Linkage disequilibrium across chromosomes . . . . .	31
3.2	Linkage disequilibrium across chromosomes; continuation . . . . .	32
3.3	Linkage disequilibrium against distance between SNP pairs within 1 Mbp . . . . .	33
3.4	Linkage disequilibrium against distance between SNP pairs within 1 Mbp; continuation . . . . .	34
3.5	Block's length . . . . .	35
3.6	Block's length; continuation . . . . .	36
3.7	Reduction in tagSNP subsets . . . . .	37
3.8	Reduction in different groups . . . . .	38
3.9	Percent of reduction in sub-population compared to the reference pop- ulation . . . . .	40
3.10	Percent of the same tagSNP between sub-populations . . . . .	41
3.11	Tagging validation . . . . .	42

4.1 LD in human and cattle populations . . . . . 46

---

## List of Tables

1.1	Single Nucleotide Polymorphism in the DNA strands of two individuals	6
1.2	Minor allele frequencies of SNPs in different populations . . . . .	7
1.3	Sources of SNPs for the BovineSNP50 assay . . . . .	8
2.1	Distribution of SNPs and genes across the genome . . . . .	17
2.2	The description of animal subsets . . . . .	27
2.3	The percentage of common individuals between groups . . . . .	27
3.1	Summary of LD blocks . . . . .	30
B.1	TagSNP subsets . . . . .	57
B.2	The reduction in each subpopulation . . . . .	58
B.3	The percent of identical tagSNPs . . . . .	59