Universität für Bodenkultur
Department für Nachhaltige Agrarsysteme
Institut für Nutztierwissenschaften
Arbeitsgruppe Tierzucht

Positioning of 15,036 Single Nucleotide Polymorphism (SNP) Loci using Order algorithms and estimates of pair wise linkage disequilibrium (LD) in Australian dairy cattle

Diplomarbeit

vorgelegt von

Markus Neuditschko

Beurteiler:

Univ. Prof. DI Dr. Johann Sölkner

Wien, im Mai 2008

"The beginning of all science is the amazing realization that all things are, as they are."

(Aristoteles)

DECLARATION

I hereby declare that I am the sole author of this thesis, entitled: **Positioning of 15,036 Single Nucleotide Polymorphism (SNP) Loci using order algorithms and estimates of pair wise linkage disequilibrium (LD) in Australian dairy cattle**, and submitted as partial fulfillment for the Master degree of Agriculture at the University of Natural Resources and Applied Life Sciences, Vienna

This body of work is the result of my own research studies, which I have done at the University of Sydney, except where otherwise acknowledged, correctly and completely.

Markus Neuditschko
May, 2008

SUMMARY

Positioning of 15,036 Single Nucleotide Polymorphism (SNP) Loci using Order algorithms and estimates of pair wise linkage disequilibrium (LD) in Australian dairy cattle:

The aim of this study was to develop a new procedure for determining Locus Order solely from pair-wise estimates of linkage DisEquilibrium – a so-called LODE map – for a high density SNP panel in dairy cattle.

Several previous studies concluded that genetic algorithms and estimates of linkage disequilibrium (LD) in livestock populations could possibly be used to position single nucleotide polymorphisms (SNPs) on an existing genome map (Goddard *et al.* 2005; Miller *et al.* 2006). The proposed formulation of genetic mapping based on LD can be considered as a version of a traveling salesperson problem (TSP) – i.e. many solutions need to be optimized to find an optimal outcome.

For the first investigations on BTA6 compromising 396 SNPs, three different order algorithms were applied, providing optimal solutions for the TSP: on our bovine sample HOPACH (Laan & Pollard 2003), Fast Optimal Leaf order (Joseph *et al.* 2001) and Sorting Points into Neighborhood (**SPIN**) (Tsafrir *et al.* 2005). The results have been discussed and compared in depth. The unsupervised novel approach of SPIN was chosen as the most suitable to create a final order of SNPs on single chromosomes and to align unknown SNPs on the current bovine map. This method has shown that Lewontin's D prime (D') is an effective parameter for providing the framework of locus order, and correlation coefficient r^2 is useful for positioning un-aligned loci. (These are two parameters most commonly used to quantify LD.)

With the current LODE map approach, the use of SPIN made it possible to determine locus order within 29 independent chromosomes, with the orders being highly correlated with the Btau3.1 map (r = 0.947) and a consensus map (r = 0.956).

In addition this study enabled separation of a set of markers from 5 chromosomes into single framework maps, and the alignment of 160 unassigned SNPs for the very first time. To investigate the generality of this current approach, the procedure was applied to murine and human samples (Fung *et al.* 2006; Valdar *et al.* 2006). In these cases, the extent of LD was either too high (murine) or too low (human) to provide useful results, indicating that the current procedure is obviously dependent on the nature and extent of LD in the population sampled.

The results of this study have made marker ordering based on LD an exciting reality. With this awareness LD mapping is now becoming an independent mapping tool, which can be used to reconsider SNP positions in current bovine maps and increase the numbers of SNPs positioned in the bovine genome.

ZUSAMMENFASSUNG

Positionierung von 15.036 Single Nucleotide Polymorphism (SNP) Loci basierend auf Schätzwerten des Gametenphasenungleichgewichtes (GPU) mit Hilfe von Ordnungs-Algorithmen in Australischen Milchrindern:

Ziel dieser Diplomarbeit war es, mit mathematischen Methoden die chromosomale Position von genetischen Markern basierend auf Schätzwerten des paarweisen Gametenphasenungleichgewichtes (GPU), von Genorten im Rinder-Genom zu bestimmen. Für die Untersuchung standen 15.036 single nucloetide polymorphism (SNP) Marker zur Verfügung, die bei 1.546 australischen Holstein-Friesian Stieren genotypisiert wurden.

Erste Ansätze und Ergebnisse bezüglich dieser neuen Methode der Markerpositionierung in Nutztieren wurden von Miller und Hayes (2006) am 8. Weltkongress für Genetik in Nutztierpopulationen in Brasilien präsentiert. Generell werden derartige Methoden in der Genetik mit einem komplexen und rechenintensiven Problem assoziiert. Die Suche einer optimalen Anordnung der SNP-Marker aufgrund von paarweisen Ähnlichkeiten bzw. Distanzen lässt sich eindrucksvoll mit einer Version des so genannten „Traveling Salesperson Problem“ (TSP) darstellen.

In dieser Arbeit wurden für die optimale Lösung des TSP die Methoden HOPACH (Laan & Pollard 2003), Fast Optimal Leaf Order (Joseph *et al.* 2001) und Sorting Points into Neighborhood (**SPIN**) (Tsafrir *et al.* 2005) angewendet. Ein Vergleich zwischen den Positionen aus der aktuellen Bovine Karte (Btau3.1) und den einzelnen Positionsergebnisse der Algorithmen zeigte, dass nur die Methode SPIN für die Positionierung der verfügbaren genetischen Markern erfolgreich angewendet werden konnte. In diesem Zusammenhang haben erste Ergebnisse gezeigt, dass Lewontin's D prime (D’) nützlich ist, um eine erste Anordnung von SNPs zu generieren. Der Korrelationskoeffizient r^2 konnte erfolgreich angewendet werden, um nicht katalogisierte SNPs auf bestehenden Gen-Karten zu positionieren und in weiterer

Folge deren physikalische bzw. genetische Position zu berechnen. Die Korrelationskoeffizienten D' und r^2 werden generell verwendet um das Ausmaß von GPU in Genomen, Chromosomen und zwischen einzelnen Markern zu beschreiben.

Mit der, in dieser Arbeit ausgearbeiteten Methode für die Positionierung von SNPs basierend auf GPU Schätzwerten ist es gelungen, hochqualitative Markerordnungen auf 29 von 30 Chromosomen zu berechnen. Die entwickelten genetischen Karten stimmten in hohem Masse mit Positionen aus der Bovine Map 3.1 ($r = 0.947$) bzw. mit Positionen aus der aktuellen Consensus Map Strategie (CRC) ($r = 0.956$) überein. In weiterer Folge konnten aus einem zufällig permutierten SNP Datensatz, welcher die Chromosomen BTA1, BTA2, BTA14, BTA28 und BTA29 beinhaltete, die SNPs erneut den einzelnen „Stamm" Chromosomen zugeordnet werden. Durch diese neue Methode der Markerplatzierung konnte die physikalische Position von 160 SNPs zum ersten Mal berechnet werden. Eine Applikation unserer Methode an einem Menschen- und Mäusebeispiel ergab keinen nennenswerten Erfolg (Fung *et al.* 2006; Valdar *et al.* 2006). In dieser Hinsicht wurde festgestellt, dass die aktuelle Methode offensichtlich von der Natur und dem Ausmaß des GPU in verschiedenen Säugetier Genomen abhängig ist.

Die Ergebnisse dieser Arbeit lassen den Schluss zu, dass unter Verwendung von GPU Schätzwerten neue unabhängige Informationen im Bereich der Positionierung von genetischen Markern liefern können. Durch diese Erkenntnis können falsch eingeordnete Marker lokalisiert werden und zusätzliche, kürzlich entdeckte Marker auf bestehenden Gen-Karten positioniert werden.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I want to thank the University of Sydney for giving me the permission and providing the resources to under take this project. Furthermore I thank the Dairy CRC for providing me with data to work with.

In particular I would like to thank the Australian ReproGen team (Frank Nicholas, Herman Raadsma, Mehar Khatkar and Matthew Hobbs), without whose help this project would not have been achieved. Special thanks also to Doctor Thomas Druml for his continuing support, advice and help throughout my stay in Camden.

Special mention must go to Doctor Professor Johann Sölkner for his supervision throughout the whole project. His patience, time, guidance and effort were unsurpassed and are greatly appreciated.

Finally, thanks must go to my family and fellow students I met in Camden, for eternizing the time in Australia with their encouragement and support.

GLOSSARY

Association mapping	Gene localization by linkage disequilibrium without cloning.
Bacterial artificial chromosome (BAC)	A type of DNA vector which can have inserts of approximately 100kb
BLAD	Bovine leukocyte adhesion deficiency
CentiMorgan (cM)	A unit for measuring genetic distance A Morgan is 100 cMs. A cM is approximately equivalent to a 1 % recombination value if (double) high levels of crossover are ignored.
Chromosome	(chroma = color, soma = body), part of the nucleus and carrier of DNA.
Comparative mapping	A genetic mapping strategy for transferring genomic information across species, based on genome homology among the species.
CVM	Complex vertebral malformation
Gametic disequilibrium	Linkage disequilibrium
Genetic maps	Maps specifying distance in crossover counts (Linkage maps) or LD units
Haplotype	Set of closely linked genetic markers present on one chromosome, which trend to be inherited together.

Malecot parameters	Parameters (M, L, ε) predicting linkage dis-Equilibrium among m markers in a physical map.
MAF (minor allele frequency)	The lowest allele frequency at a locus that is observed in a population.
Principal Components Analysis (PCA)	Linear dimensionality reduction technique that seeks to identify a small number of "dimensions" or "components" that capture most of the relevant structure in the data.
Quantitative trait loci (QTL)	Genes controlling quantitative traits
Physical map	Map specifying distance in the DNA sequence, ideally measured in bp. Less reliable physical maps provided by chromosome bands and breakage in radiation hybrids.
Recombination fraction	Ratio between recombinant gametes produced by meiotic events. It is commonly estimated by maximizing likelihood functions which are built using the observed genotypic frequencies in mapping populations and the expected genotypic frequencies as function of recombination fractions.
SNP	Single Nucleotide Polymorphism kind of genetic Marker
Traveling salesperson problem	The problem of finding the shortest cyclical itinerary for a traveling salesman who must visit each city, optimally once, given a symmetric matrix of distances among a set of cities.

TABLE OF CONTENTS

Declaration 3
Summary 4
Zusammenfassung 6
Acknowledgements 8
Glossary 9
Table of Contents 11

1. INTRODUCTION 13

2. LITERATURE REVIEW 15

2.1. GENETIC MARKERS 15
2.2. LINKAGE DISEQUILIBRIUM 16
2.2.1 MEASUREMENT OF DISEQUILIBRIUM 16
2.2.2 LINKAGE DISEQUILIBRIUM AND GENETIC DISTANCE 18
2.2.3 LINKAGE DISEQUILIBRIUM IN FOUR POPULATION SAMPLES 19
2.3. THE TRAVELING SALESPERSON PROBLEM 21
2.3.1 DEFINITION 21
2.3.2 APPLICATIONS AND SOLUTIONS 22
2.4. MAP BUILDING 22
2.4.1 PHYSICAL MAPPING 23
2.4.2 GENETIC MAPPING 24
2.4.2.1 Linkage mapping 24
2.4.2.1.1 Linkage Grouping 24
2.4.2.1.2 Linkage Grouping Criteria 25
2.4.2.1.3 LOCUS ordering 25
2.4.2.2 LD mapping 26
2.4.2.3 Comparative mapping 28
2.5. RESUME 29

3. MATERIALS AND METHODS 31

3.1. BACKGROUND MATERIAL 31
3.1.1 SNP GENOTYPES 31
3.1.2 PHYSICAL MAPPING 31
3.1.3 INTEGRATED MAPPING 32
3.1.4 LINKAGE DISEQUILIBRIUM DETERMINATION 33
3.2. METHODS 33
3.2.1 CRITERIA TO ORDER AND POSITION SNPS USING LD INFORMATION 33
3.2.2 APPLIED ORDERING ALGORITHMS 34
3.2.2.1 HOPACH 34
3.2.2.2 Fast optimal leaf ordering 34
3.2.2.3 **SPIN** (Sorting Points into Neighborhoods) 35
3.2.2.3.1 Side to Side (STS) Algorithm 36
3.2.2.3.2 Neighborhood algorithm 37
3.2.3 DATA SUBSETS 38

3.2.3.1 Test batch 1 – Full bovine LODE Map 39
3.2.3.1.1 Fisher r-to-z transformation 39
3.2.3.2 Test batch 2 – Alignment of high quality known SNPs as unknown 39
3.2.3.3 Test batch 3 – Alignment of low quality and problem SNPs 40
3.2.3.4 Test batch 4 – Alignment of current unaligned SNPs 40
3.2.4 THE CALCULATION OF SNP POSITION 40

4. RESULTS 41

4.1. ORDERING ALGORITHM COMPARISON 41
4.2. METHOD TESTS 41
4.3. MINIMUM SAMPLE SIZE 43
4.4. WHOLE GENOME ANALYSES 44
4.5. ALIGNMENT PROCEDURE 45
4.6. TEST BATCH 1 – FULL BOVINE LODE MAP 46
4.6.1 THREE MAP COMPARISON 48
4.7. TEST BATCH 2 – ALIGNMENT OF HIGH QUALITY KNOWN SNPS AS UNKNOWN .. 51
4.8. TEST BATCH 3 – ALIGNMENT OF LOW QUALITY AND PROBLEM SNPS 52
4.9. TEST BATCH 4 – ALIGNMENT OF CURRENT UNALIGNED SNPS 54

5. DISCUSSION 55

5.1. GENERAL FEATURES OF THE LODE MAP 55
5.2. SPECIFIC FEATURES OF THE LODE MAP 58
5.2.1 ORDER PROCEDURE 58
5.2.2 ALIGNMENT PROCEDURE 60
5.3. THE LODE MAP STRATEGY 64

6. CONCLUSION 65

6.1. GENERAL STATEMENTS ON THE UTILITY OF A BOVINE LODE MAP 65

7. REFERENCES 66

8. APPENDIX 72

8.1. WHOLE MURINE GENOME RESULTS 72
8.2. POSITIONS OF THE 160 ALIGNED SNPS 74

1. INTRODUCTION

Genetic markers (SNPs) are measurable patterns in DNA that may correlate with particular traits in animals or plants. Linkage disequilibrium (LD) is also known as allelic association, occurs when two alleles at adjacent loci (SNP) are found together in a chromosome more often as expected in a population (Trapper *et al.* 2003). This genetic phenomenon aids our association between these markers and quantitative trait loci (QTL). Hence marker assisted selection exploiting linkage disequilibrium would make it possible to detect important QTL's as well as undesirable genes such as bovine leukocyte adhesion deficiency (BLAD) and complex vertebral malformation (CVM) (Hayes *et al.* 2006).

However for a functional genome-wide SNP map to be composed at least 500,000 SNPs will be required (Kruglyak 1999) in human and likely up to 300,000 SNPs in cattle (Herman Radsmaa, personal communication). Fortunately the cost of SNP discovery and genotyping is rapidly decreasing, allowing hundreds or even thousands of individuals to be genotyped for hundreds of SNPs (Käller *et al.* 2007).

As a result of a current Genetic Marker Project study at the University of Sydney (http://www.vetsci.usyd.edu.au/reprogen/) a data set of 15,036 Single Nucleotide Polymorphism (SNP) markers were identified through high genotyping (http://www.affymetrix.com) in Holstein Friesian Bulls (n = 1,546).

Hence the Innovative Dairy Products Research Center (CRC) has developed a consensus mapping strategy that combines the four major independent bovine maps (BAC, USDA MARC, ILTX3, SIAG and BovGen). Using this new strategy it was possible to reduce the number of unmapped SNPs compared to the current bovine map. However the positions of 300 SNPs were still unknown. In this case LD between markers could possibly be used to infer the position of the currently unmapped markers (Goddard *et al.* 2005) as markers closely co-located within the genome are expected to show a higher degree of LD.

The primary objective of this study was to find mathematical/statistical methods (algorithms) to create a new type of genetic map based solely on pair-wise estimates of linkage DisEquilibrium.

The two issues (hypotheses) addressed in this study are:

- The use of pair wise LD between genetic markers to generate a final order of SNPs on an existing bovine map (Btau3.1);
- And the allocation of unknown SNPs based on pair-wise LD.

2. LITERATURE REVIEW

2.1. Genetic Markers

To understand the history and evolution of populations it is usually necessary to study a large number of polymorphisms (Cavalli-Sforza 1998). Through molecular revolution over the last few decades, a lot of techniques have been developed using genetic markers. At the first stage of research almost all markers identified have been protein polymorphisms, and only a few hundred (Nei & Roychodhury 1988) were previously known. These markers are also known as "classical polymorphisms" to distinguish them from those obtained by direct DNA analyses. The use of DNA segments to analyze genetic polymorphisms has resulted in the identification of a great number of markers and genetic polymorphisms, which is of great benefit as a lot of markers are needed to study specific problems.

Commonly considered DNA markers are single nucleotide polymorphisms (SNPs or "snips"). A SNP is a small (single pair) genetic change, or variation, that can occur within an individuals DNA sequence. This property of SNPs allows us to make associations between marker allele studies and diseases. It has been hypothesized that these studies are our most powerful method identifying genes that cause common diseases such as heart disease, cancer, diabetes and psychiatric illness. SNPs are a good choice of marker for these studies because of their low mutation rate, high incidence throughout the genome and bi-allelic nature (making them amenable to automated detection techniques) (Dawson 1999).

The potential advantage of LD mapping over conventional linkage analysis performed within families lies in the use of 'historical' recombinants, thereby increasing resolution (Hästbacka *et al.* 1992; Talbot *et al.* 1999) and power of association studies. Elementary for such association studies based up on LD mapping are the number of SNP and their linkage disequilibrium values.

2.2. LINKAGE DISEQUILIBRIUM

Linkage disequilibrium (LD) is a measure of the amount of recombination, after numerous generations of random mating that has occurred between two regions of the genome (loci). Given that recombination occurs throughout the genome, and is proportional to distance between loci, it is therefore expected that LD values are indicative of how far apart the loci are. Consider two SNP markers (A and B), for example – each having two alleles with equal (50%) frequency. A and a are the alleles at marker locus A, and B and b are the alleles at the marker locus B. If the two SNPs are at the opposite ends of the chromosome, then throughout many generations in which they have both existed there will be recombination between the two loci, resulting in equal numbers of chromosomes carrying the four possible haplotypes (AB, Ab, aB, ab). However, if two loci are close together the recombination between them will be less frequent, resulting in SNPs on associated alleles baring similarity. This allele association between the SNP markers is also a statistical association of sequence variants of different positions along the chromosomes as they occur in gametes, called LD. The significance of LD is influenced by two attributes of LD, that should be kept distinct are the statistical significance of LD, which depends on the sample size, and the magnitude of LD (Weiss & Clark 2002).

2.2.1 Measurement of Disequilibrium

There are several different measures for disequilibrium (Hedrick 1987; Lewnontin 1988) in a two-locus model. If there are two loci, A and B, each with two alleles (A and a; B and b), the allelic frequencies are:

p_A, $1 - p_A$	for	loci A (A, a)
p_B, $1 - p_B$	for	loci B (B, b)

If the two loci are independent and each of the loci is in Hardy-Weinberg equilibrium, then the expected gametic frequencies are:

$$p_{AB} = p_A p_B$$
$$p_{Ab} = p_A (1 - p_B)$$
$$p_{aB} = (1 - p_A) p_B$$
$$p_{ab} = (1 - p_A) (1 - p_B)$$

Disequilibrium is due to many phenomena. For example mutation of an existing gene or recent population admixture can produce disequilibrium. However, the most influential cause of disequilibrium extent from generation to generation is genetic linkage. Departure form equilibrium is commonly measured by a two-locus coefficient of gametic disequilibrium (D_{AB}). This coefficient is decreasing by a factor of (1 – r) each generation of random mating, if the two loci are linked with a recombination fraction of r. Thus a high recombination fraction causes low LD in a population; the genetic linkage in the population is low and vice versa respectively.

$$D_{AB} = p_{AB}p_{ab} - p_{Ab}p_{aB}$$
$$D_{AB}^{t+1} = (1 - r) D_{AB}^{t}$$

The gametic disequilibrium coefficient can be positive or negative. The general range of D_{AB} depends on the allelic frequencies. A measure that attempts to avoid this dependence on allelic frequencies is Lewontin's |D'| which is:

$|D'| = D/D_{max}$ (D_{max} is the lesser of $p_A p_B$ if D is positive or $p_A p_B$ or $p_a p_b$ if D is negative).

Another measure of LD is the square of the correlation coefficient between the A and B loci:

$$r^2 = D^2/p_A p_a p_B p_b$$

R^2 measures statistical association between the possible haplotypes, an r^2 value of 1 indicates that only two haplotypes are present.

The interrelated parameters D' and r^2 are the most frequently used measures of linkage disequilibrium (Pritchard & Przeworski 2001). The decay of D' with distance is much slower than that of r^2. D' and r^2 values range from 0 to 1, with higher values indicating closer vicinity. Some other measures of pair wise LD which have been proposed are $\Delta=r$ (Hill & Weir 1994) and Ψ (Edwards 1963).

2.2.2 Linkage Disequilibrium and Genetic distance

The measure of genetic distance allows one to quantify genetic relationships between two samples, like pair-wise LD. It is used to describe the proportion of genetic elements (alleles, genes, gametes and genotypes) that the two samples do not share. Depending on the similarities (S) of the samples the genetic distance can vary from 0 to 1. There is a genetic distance of 1, when the two samples have no genetic element in common.

Depending on the nature of a data set genetic distances can be calculated in three different ways and used for 2 different types of genetic models (Figure 1) (Cornell & IPGRI 2003).

1. $D = 1 - S$, known as linear distance, because it assumes that the relationship with similarity (S) is linear.
2. $D = \sqrt{(1 - S)}$, known as quadratic distance, the similarity relationship follows a quadratic function, so that, to make it linear, the square root must be calculated.
3. $D = \sqrt{(1 - S^2)}$, describes a circular distance.

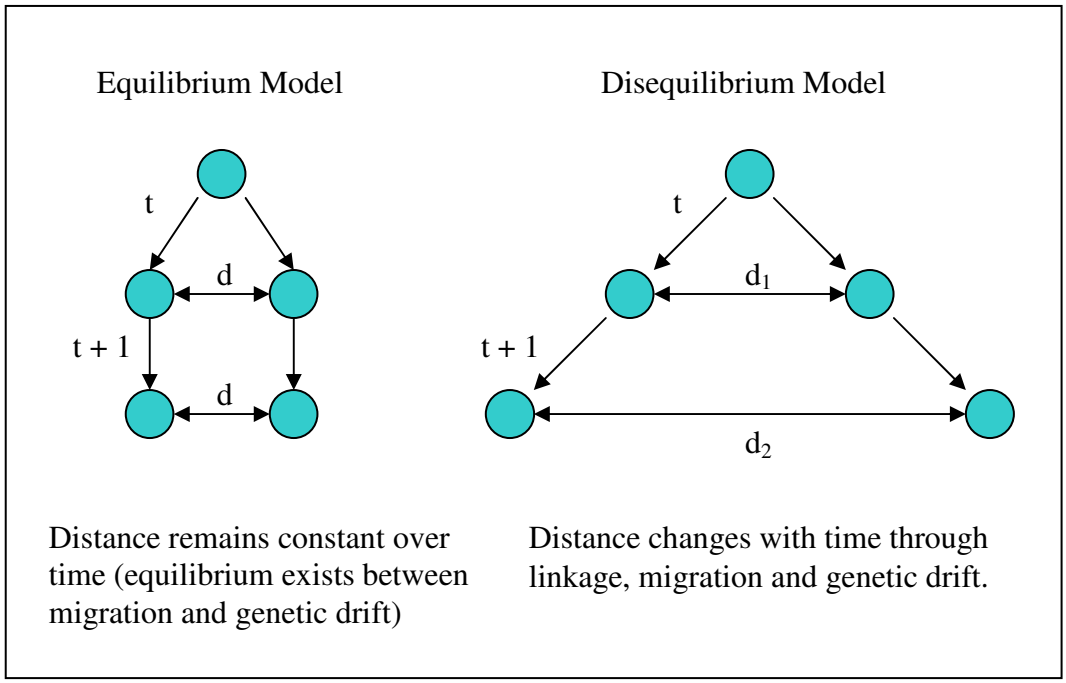


Figure 1 Distance Models

D' and r^2 are linear relationships, thus these values can be easily transformed into a genetic distance by subtracting them from 1. While the distance will be influenced by linkage, migration, genetic drift and other circumstances, these are accounted for D' and r^2 values.

2.2.3 Linkage disequilibrium in four population samples

Linkage disequilibrium (LD) varies enormously among populations, due to many factors, including genetic drift, population growth, admixture, population structure and mutation rates and gene conversion (Weeks & Lathrop 1995; Ardlie *et al.* 2002). For this reason, the human International HapMap project was developed to describe disequilibrium patterns in diverse ethnic groups. The significance of LD is low for the whole human population due to the extent of LD. Thus LD studies have attracted commercial interest in genetically isolated populations. For example DeCode Genetics is studying the population of Iceland and its extensive genealogy to identify genes for common diseases, as this young population's recombination has had less time to whittle down LD compared to other populations (Figure 2 HSA15, HSA22). Compared to productive livestock, where selection and artificial insemination is

widespread, the LD of human isolated population is still much lower (Figure 2 HSA15, HSA22, BTA1).

The extent of LD in livestock is expected to be higher than in humans, since the factors mentioned above are more extreme in livestock populations (Nsengimana & Baret 2004). The evidence supporting this expectation is constantly increasing: High LD levels have been found in dairy cattle (Khatkar *et al.* 2006) sheep (Farnir *et al.* 2000), pigs (Nsengimana *et al.* 2004) and horses (Tozaki *et al.* 2005).

The mouse data described in Figure 2 is derived from an artificial population, which has been cultured to get fine-mapped quantitative trait loci (QTLs) for complex traits in mice (Valdar *et al.* 2006). Therefore this mouse population was derived form 8 inbred lines. As a result of the very small population size the mouse population data shows a higher extent of LD compared to the cattle and human data sets. (Figure 2 HSA22, BTA1 MMU9).

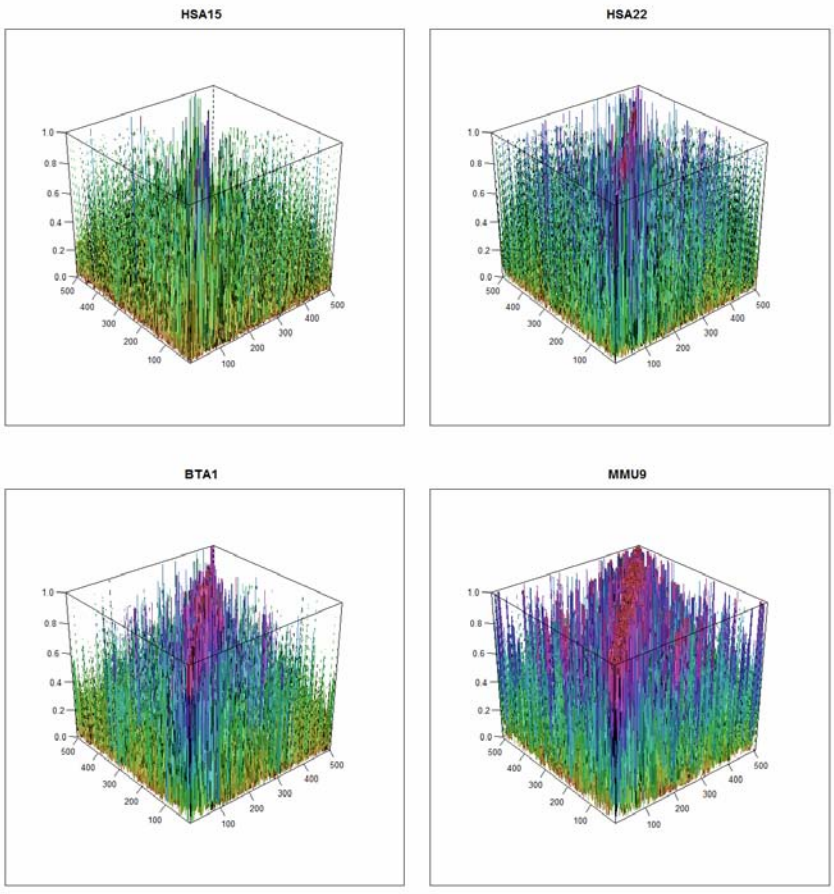


Figure 2 D' heat maps of 4 different populations, American humans (HSA15), Finnish sub isolate (HSA22), Australian Dairy Cattle (BTA1) and Mouse (MMU9)

Human SNP genotyped on HSA15 in 267 Parkinson's disease patients are catalogued on the SNP Database at the NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds/).

SNP data genotyped on HSA22 in 200 Finnish sub isolate on chromosome 15 (Varilo *et al.* 2003).

SNPs genotyped on BTA1 in 1,546 bulls for the BTA1 are from the current Marker Project at the University of Sydney (http://www.vetsci.usyd.edu.au/reprogen/).

The murine set consists of 13,459 SNP of 2,002 heterogeneous stock mice available online at (http://gscan.well.ox.ac.uk).

2.3. The Traveling Salesperson Problem

2.3.1 Definition

A generic description of the symmetric TSP is: Given a list of cities and costs c_{ij} for traveling between all the pairs of cities, the TSP involves specifying a minimum-cost tour where each city must be visited once and returning to the starting point at the end of the trip (Miller & Pekny 1991). The mathematical formulation of this problem is a symmetric matrix of costs among a set of l cities denoted by the set of V = {1,2, …,l}, the formulated hypothesis is to find the shortest cyclical itinerary for the traveling salesperson who must visit each of l cities where the costs c_{ij} should be minimized. Then the problem becomes to find an order with

$$\sum_{i \varepsilon V} \sum_{i \varepsilon V} c_{ij} x_{ij}$$

Maximized, subject to

$$\begin{cases} \sum_{i \varepsilon V} x_{ij} = 1, \; j \varepsilon V \\ \sum_{i \varepsilon V} x_{ij} = 1, \; j \varepsilon V \\ \sum_{i \varepsilon V} \sum_{i \varepsilon V} x_{ij} \leq |S| - 1 \qquad for \qquad S \subset V, S \neq 0 \\ x_{ij} \varepsilon \{0,1\}, \; i, j \varepsilon V \end{cases}$$

Where $x_{ij} = 1$ if (i,j) is in the solution and $x_{ij} = 0$ otherwise. Concerning gene ordering there is one constraint (Liu 1998).

$$\sum_{i \varepsilon V} \sum_{j \varepsilon V} x_{ij} \leq |S| - 1$$

This variation of the traveling salesperson problem is also called the wandering salesperson problem (WSP), which guarantees a linear order of the loci i instead of a circle (Mester *et al.* 2004).

2.3.2 Applications and Solutions

The Traveling Salesperson Problem (TSP) is attracting the attention of mathematicians, computer scientists and practitioners of many branches. It has already been studied in many different fields including x-ray crystallography, circuit board drilling, very large scale integrated circuit fabrication, circuit board assembly and in protein conformation studies (Miller & Pekny 1991).

The main problem with investigating the TSP is finding an exact solution, in an appropriate time when dealing with extensive data sets. In the last century a lot of algorithms have been proposed for exact solution of TSP, one of them is the exact *branch and bound* algorithm. Such algorithms like *branch and bound* have been successfully implemented for the construction of radiaton hybrid maps (Ben-Dor *et al.* 2000), linkage maps (Tan & Fu 2006), and for the integration of maps of the same or of different types (Mester *et al.* 2006; Faraut *et al.* 2007).

2.4. Map Building

Generally there are no differences between gene maps and road maps. The aim of a gene map is to get information about distances of loci by the determination of physical or genetic positions of markers. Knowing the position of genetic markers enables the detection of DNA fragments linked with a gene. As a result of this classification of genomic DNA fragments disease genes (e. g. BLAD or CVM) and production genes (e. g. milk yield) can be investigated.

Different technologies have been developed for gene mapping. Nowadays the most commonly used methods are:

- Somatic cell hybrids mapping (physical mapping)
- Fluorescence in situ hybridization of chromosomes (FISH) (physical mapping)
- Analyses of recombination frequencies in families (genetic mapping)
- Gene map comparisons between different species (comparative mapping)

2.4.1 Physical Mapping

Physical mapping depends on a large collection of cloned DNA fragments. For the creation of these DNA segments the polymerase chain reaction (PCR) is used, which makes it possible to rapidly generate a very large number of copies of a specific region of DNA. For the physical ordering of the created genomic fragments several methods are used like fluorescence in situ hybridization (FISH) (Montanaro *et al.* 1991; Korenberg *et al.* 1992), somatic cell hybrids (Cox *et al.* 1990) or fingerprinting methods (Craig *et al.* 1990; Stallings *et al.* 1990).

The goal of physical mapping is to create a true order of genetic landmarks (cloned DNA fragments), generated through standard positional cloning, using different physical mapping strategies, which are based on standard map construction algorithms like simulated annealing (Cuticchia *et al.* 1992; Mott *et al.* 1993), back track (Christof *et al.* 1997), resample techniques (Wang *et al.* 1994) and clustering strategies (Mayraz & Shamir 1999; Heber *et al.* 2000). With all these methods the map distance between loci can be accurately calculated.

Another strategy is taking advantage of the existence of large insert libraries of cloned DNA fragments according to their position in the genome. Such library information's for livestock species can be found in YAC (yeast artificial chromosome) (Broom & Hill 1994; Alexander *et al.* 1997) and BAC (bacterial artificial chromosome) (Morton 1955; Cai *et al.* 1995; Buitkamp *et al.* 2000).

The aforementioned characteristics of the physical map approach make it a powerful tool for the localization and isolation of genes, for studying the organization and evolution of genomes as a preparatory step for efficient sequencing.

2.4.2 Genetic Mapping

2.4.2.1 Linkage mapping

Linkage mapping confirms a specific linear arrangement of a group of genes and, or markers. It also determines which chromosome contains the genes as well as markers and indicates their respective location (precision dependent on scale of the linkage study). Locus positioning is done directly by determining the frequency of recombinants as a result of meiotic events between sets of loci. The distance between two loci is described by the centimorgan instead of the recombination frequency, e.g. a recombination frequency of 0.01 (1 percent) is defined as 1 centimorgan (cM) in honor of Thomas Hunt Morgan.

Genetic linkage analyses have already been studied in almost all livestock species including ovine (Crawford *et al.* 1995), caprine (Vaiman *et al.* 1996), bovine (Bishop *et al.* 1994) and equine (Swinbrune *et al.* 2000).

2.4.2.1.1 Linkage Grouping

Linkage grouping is the first of two steps in a locus ordering process. Linkage grouping is the partitioning of loci into linkage groups based on their linkage relationship. Chromosomes are normally the basis of linkage groups, which can be defined biologically as a group of genes with their loci located on the same chromosome or statistically as a group of loci inherited together according to statistical criteria. The linear order of the loci in a linkage group or loci locations on chromosomes is created through locus ordering algorithms.

2.4.2.1.2 Linkage Grouping Criteria

The main statistical criterion for a pair of loci (A and B) is a two point recombination fraction between recombinant gametes produced by meiotic events. It is commonly estimated by maximizing likelihood functions which are built using the observed genotypic frequencies in mapping populations and the expected genotypic frequencies as functions of recombination fractions, a lod score (similarity value between pair of loci like D' and r^2), (Morton 1955) and a significant P-value. Recombination fractions and significant P-values and lod score referred to as linkage grouping criteria and are used to determine whether loci A and B belong to the same linkage group (chromosome) or not. For example:

- if { [$\theta_{ij} \leq c$] and [$p_{ij} \leq b$] },
- if { [$\theta_{ij} \leq c$] or [$p_{ij} \leq b$] },

- { [$\theta_{ij} \leq c$] and [$z_{ij} \geq a$] } or
- { [$\theta_{ij} \leq c$] or [$z_{ij} \geq a$] }
- θ_{ij}, z_{ij} and p_{ij} denote a two-point recombination fraction

- c is the maximum recombination fraction value to be declared a linkage
- b is the maximum significant P-value for declaring a linkage
- a is the minimum lod score value to declare a linkage

If one of these criteria are fulfilled then the loci i and j belong to the same linkage group.

2.4.2.1.3 LOCUS ordering

Locus ordering algorithms create a final order of loci in a linkage group. Given the n!/2 possible orderings of n loci, ordering based on pair-wise distances is a non-deterministic polynomial problem (e. g. 20 loci in a linkage group creates 1,22 x 10^{18} of marker orders). Locus ordering algorithms search for the best locus ordering, yet these procedures are comparable with finding a needle in a haystack. Thus many

variations of ordering algorithms have been proposed using information based on pair-wise linkage data such as *seriation* (Buetow & Chakravarti 1987), *minimum sum of adjacent recombination fractions* (SARF) (Falk 1989), *minimum product of adjacent recombination fractions* (PARF) (Wilson 1988) and *maximum sum of adjacent lod scores* (SALOD) (Weeks & Lange 1987). All these ordering algorithms are aimed at minimizing the sum of adjacent recombination frequencies to create a minimum distance map and is similar to solving the shortest tour in the "traveling salesperson problem" (Miller & Pekny 1991). Solving the loci ordering problem can therefore be done by using algorithms for the TSP.

2.4.2.2 LD mapping

Linkage analyses in livestock populations as well as in human are limited, due to the small number of individuals in each study and relatively low number of recombination events, yielding relatively low resolution of loci position. To overcome this limitation, association mapping is used for gene localization. This mapping tool uses linkage disequilibrium as an association of genes at the population level, without cloning and uses historical recombination information over many generations. In contrast to linkage maps, LD maps determine distance by LD instead of recombination. To determination this distance the association probability (ρ) has to be predicted by the Malecot equation.

This ρ value depends on the allelic frequencies between a pair of diallelic loci (A and B). When ρ = 0, there is linkage equilibrium, When ρ = 1 there is complete disequilibrium (Morton *et al.* 2001).

In an extreme example of LD, where the founder haplotype frequencies are a mixture of LD with probability ρ and complementary frequency with LD = 0, the Malceot equation becomes:

$$\rho_0 \begin{bmatrix} A & 0 \\ A-B & 1-A \end{bmatrix} + (1-\rho_0) \begin{bmatrix} AB & A(1-B) \\ (1-A)B & (1-A)(1-B) \end{bmatrix} \qquad (1)$$

Where the frequency of the rarest allele is A and the frequency of the associated allele is B then ρ_0 is defined as the association probability in founders, the decay of ρ in t generations is due to mutation rate (v), effective population size (N) and recombination $e^{\theta t} \rightarrow 0$ (L).

$$\rho_t = (1 - L)\, M e^{-\theta t} + L \quad where \;\; M = (\rho_0 - L)\, e^{(v + 1/2N)t} / (1 - L) \qquad (2)$$

In cases of large t values usually the prediction for ρ_t can not be calculated over L. Therefore Haldane's mapping function is applied to improve the estimate of the distance (Collins *et al.* 1999). This mapping function counts cross over events (single once, double twice) to adjust the proportion of observed recombinant genotypes, thereby the recombination fraction will converted into map distance (Haldane 1919). The Malecot equation becomes:

$$\rho_t = (1 - L)\, M e^{-\varepsilon d} \quad e^{-\varepsilon d} \; is\ used\ instead\ of \; e^{-\theta t} \qquad (3)$$

- Where M is 1 for monophyletic origin and <1 otherwise
- L describes residual association at large distance
- $\varepsilon \geq 0$ depends on the number of generations during which the haplotypes have been approaching equilibrium
- $d \geq 0$ is the distance on the genetic or physical map
- εd equals the product of recombination and time

With this extension of the Malecot Equation the decay of LD can be illustrated based on physical (kb) or linkage (cM) mapping (Figure 3). Thus linkage maps can be enhanced by interpolating dense locations from LD maps and more importantly association mapping can be done using LD maps measured by $\sum\varepsilon$, $\sum\varepsilon_i d_i = 1$ defining 1 LD unit (LDU) also called a "swept radius".

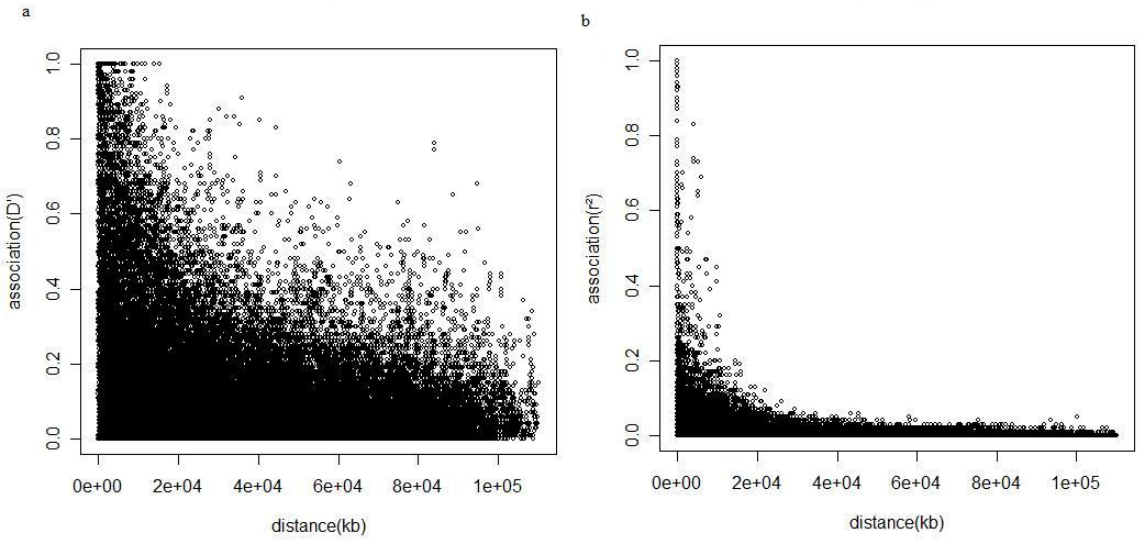


Figure 3 Predictions of ρ for BTA6 in Australian dairy cattle, with n = 1,546 and $\rho_0 = 1$, (a) Decay of ρ with D' association, (b) Decay of ρ with r^2 association.

2.4.2.3 Comparative mapping

Comparative mapping is based on Haldane's report where genes which are linked or located closely on the same chromosome are inherited together. The strategy of comparative mapping is therefore to use this conclusion and compare the arrangement of genes and DNA markers between species. This mapping strategy is especially used in species (human and livestock), where it is hard to accumulate mapping information about important genes

For example important quantitative traits in human and livestock like high blood pressure (HTN) have been intensively studied in rat and mouse models to identify the genes that contribute for this QTL (Lee *et al.* 2000) as identified on rat chromosome 2 (Jeffs *et al.* 2000). To transfer this candidate region to the human genome tools like expressed sequence tags (ESTs) and basic alignment search tool (BLAST) (Altschul *et al.* 1997) are used. A set of ESTs can be used to identify a region in the human genome containing many of the same sequences (syntenic region) found in the rat genome. In case of HTN studies chromosome 2 in rats was highly related to an area of human chromosome 1. By using BLAST, a comparative human chromosome 1 and a rat chromosome 2 was generated.

2.5. Resume

In many years of genetic science much effort has been invested in creating high resolution maps, to identify genetic regions of functional importance like QTL's. Such loci have an important biological and economical effect, but are usually hard to find by only using molecular identification tools. To overcome this limitation new mapping tools based on genetic relationships (linkage and linkage disequilibrium) have been developed to get exact loci positions. This importance of genomics to fundamental biology is shown by Lander and Schrock's (1994) paper on genetic dissection of complex traits, stating

"one can systematically discover the genes causing inherited diseases without any prior biological clue as to how they function. The method of genetic mapping, by which one compares the inheritance pattern of a trait with the inheritance patterns of chromosomal regions, allows one to find where a gene is without knowing what it is"

In this respect LD maps have already shown higher relative efficiency than physical maps concerning the description of LD (Zhang *et al.* 2002) and association mapping (Maniatis *et al.* 2004), especially in population wide studies. However current LD maps still rely on physical or linkage maps to describe the distance between loci (Morton *et al.* 2001).

To overcome the problem of creating LD maps based on prior known positions this study will act in reverse, by allocating loci positions (SNPs) based on LD relations. This new method is describing the inverted use of the Malecot equation for positioning loci on a physical map based on linkage disequilibrium (D' distances) (Figure 4). Hence, LD maps will become more independent form prior physical information.

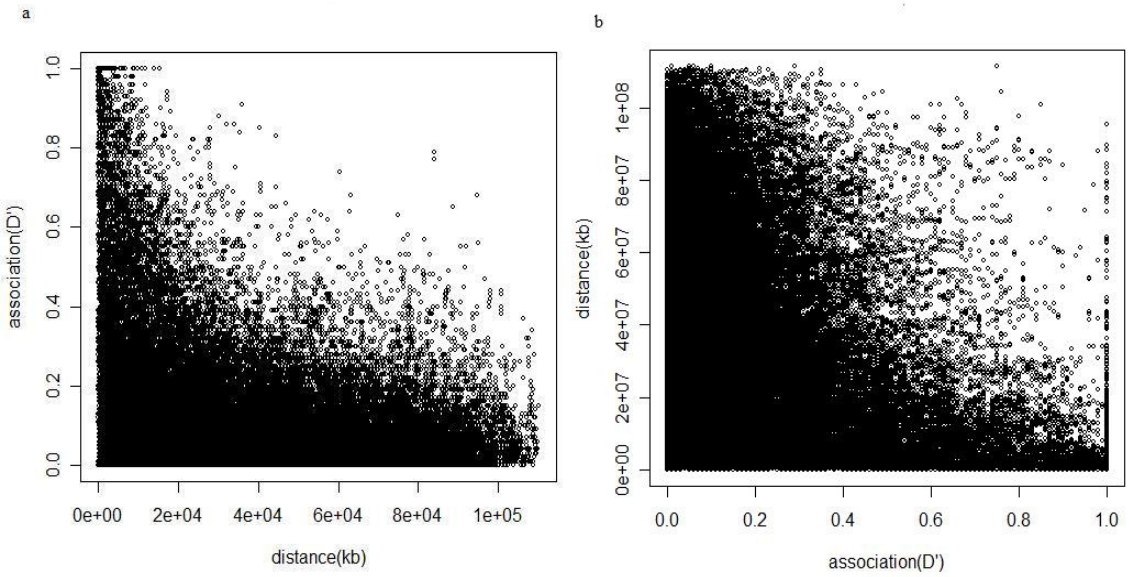


Figure 4 Decay of D' association on BTA6 in Australian dairy cattle, (a) Prediction of D' association based on physical distance (kb), (b) Prediction of physical distances based on D' association.

Using this new approach, LD maps can be created in species for which no prior mapping information is available without the need to establish a set of mapping families. Furthermore this new kind of genetic map (LODE map) is strengthening the use of linkage disequilibrium for SNP marker positioning and genome structure analysis.

3. MATERIALS AND METHODS

3.1. Background Material

3.1.1 SNP genotypes

A total of 15,036 Single Nucleotide Polymorphisms (SNPs) genotyped in 1,546 Holstein-Frisian bulls were obtained from the current bovine data set at the University of Sydney (Khatkar *et al.* 2007).

To investigate the general application of the current LODE map approach to genome-wide ordering, we calculated LD for a panel of 500 SNPs genotyped across HSA15. In addition a total of 13,459 SNPs genotyped in 2,002 heterogeneous stock mice, available online at http://gscan.well.ox.ac.uk, were used.

Human SNP are derived from the SNP Database at the NINDS Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds/). We used a data set with more than 408,803 SNP genotyped in a cohort of 267 Parkinson's disease patients and 270 neurologically normal controls (Fung *et al.* 2006), all of Caucasian origin. The control cohort consisted of individuals collected from several sites across North America (Simon-Sanchez *et al.* 2007).

3.1.2 Physical mapping

The positions of the 15,036 SNPs in the bull data set were determined on the bovine sequence assembly Btau3.1.
(ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20060815-freeze/)

3.1.3 Integrated mapping

The process of integrated mapping (also called IDPCRC0207) determines SNP positions based on a consensus between the major independent bovine maps available. These include BAC (bacterial artificial chromosome), USDA MARC (linkage), ILTX3, SIAG and BovGen (Radioactive hybrid) map. Better SNP positioning is achieved when a combination of mapping strategies are applied, compared to the current Btau3.1 assembly (Figure 5).

The advantages of such a mapping strategy are that it:

- Takes only a few days to create a consensus scaffold order for all chromosomes;
- makes use of all major independent maps;
- includes an extra 1,217 scaffolds, in addition to the 3,503 scaffolds in Btau3.1;
- agrees more closely with each of the independent maps and
- agrees more closely with sequence assemblies in other species.

Therefore both maps (Btau3.1 and Consensus) were used as reference for comparison with LODE maps.

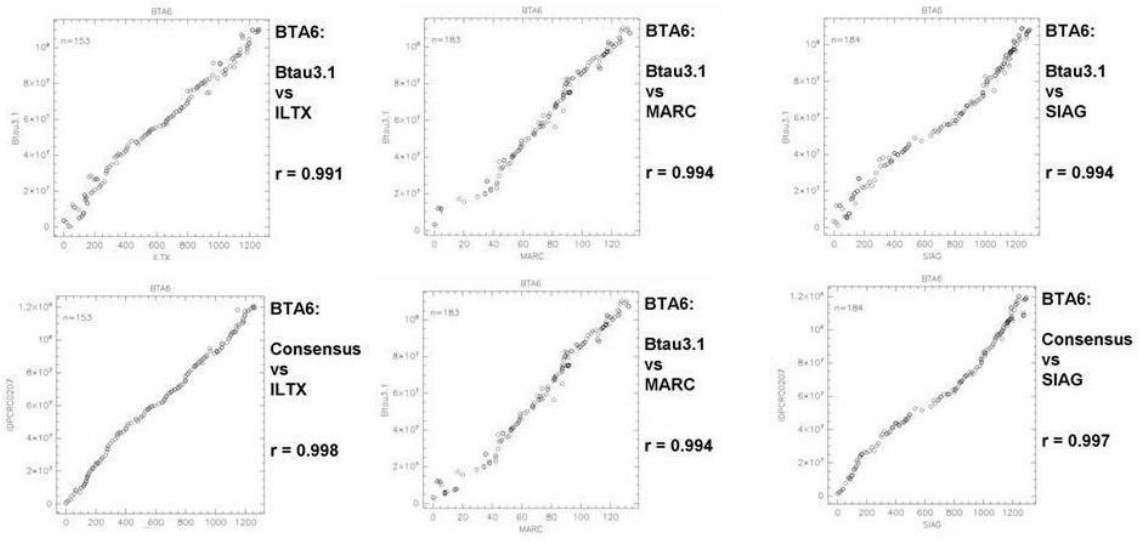


Figure 5 Comparison of locus order in three of the major separate (independent) bovine maps with Btau3.1 (first row) and consensus (second row) locus order, for BTA6

3.1.4 Linkage disequilibrium determination

Linkage disequilibrium measures, D' and r^2 of pairs of loci for the three populations (bovine, murine, human) were calculated by Mehar Kathkar using Haploview (Barrett *et al.* 2005).

3.2. Methods

3.2.1 Criteria to order and position SNPs using LD information

The possibility to position SNPs on an existing bovine map, by exploiting linkage disequilibrium, has already been shown by Miller and Hayes (2006). In this study, the most likely relative positions of a set of n SNPs were determined by minimizing:

$$\sum_{i=1}^{n} \sum_{j=1}^{i-1} d_{ij} * r_{ij}^2 \qquad [1]$$

In our study we used instead of the absolute value for the distance between $locus_i$ and $locus_j$ a genetic linear distance $(1 - S)$ with D'_{ij} and r^2_{ij} respectively as similarity values (S), for example adjacent loci *i* and *j* in our study have a d_{ij} of 0 ($S = 1$) instead of 1, as shown in the equation [1] above. Hence the SNP order and position respectively, which minimizes the outcome of the equation [2] below, is considered to be the most likely order and position respectively.

$$\sum_{i=1}^{n} \sum_{j=1}^{i-1} d_{ij} \quad with\ d_{ij} = \left(1 - S\right) \qquad [2]$$

3.2.2 Applied ordering algorithms

For determining the most likely ordering of a set of SNPs based on LD, algorithms used in gene expression analysis like HOPACH- see below (Laan & Pollard 2003), Fast optimal leaf ordering for hierarchical clustering (Joseph *et al.* 2001) and Sorting Point in Neighborhood (SPIN); (Tsafrir *et al.* 2005) have been adopted.

3.2.2.1 HOPACH

Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) creates a meaningful order of elements (in this study loci) by generating a hierarchical tree of clusters. This clustering method uses a greedy, step-wise algorithm (*improveordering*) as a prior step to improve the ordering of the elements in a distance matrix by swapping the ordering of the SNPs till no further improvement in ordering can be achieved. This process is not yet optimized and as a result this function is very slow when more than 50 elements are considered (Laan & Pollard 2003).

3.2.2.2 Fast optimal leaf ordering

Fast optimal leaf ordering is a practical algorithm for the optimal linear leaf ordering of trees that were generated by hierarchical clustering, and is used extensively to analyze gene expression data. The aim of this algorithm is to find a ordering of tree leaves, that maximizes the sum of similarities of adjacent leaves (loci) in the ordering. The objective function of this algorithm is defined as follows:

$D^{\pi}(T) = \sum_{i=1}^{n-1} S(\pi_i, \pi_{i+1})$ where S is the data similarity matrix and π stands for the ordering that maximizes the possible orderings of the tree leaves (Joseph *et al.* 2001).

The clustering algorithms HOPACH and Fast optimal leaf order were considered to be optimal solutions for the SNP ordering problem, based on their finer clustering structure. Contrary to other clustering approaches both algorithms generate a useful

order of elements in the clusters (Figure 6). Hence, this approach follows a similar principle to other linkage ordering procedures.

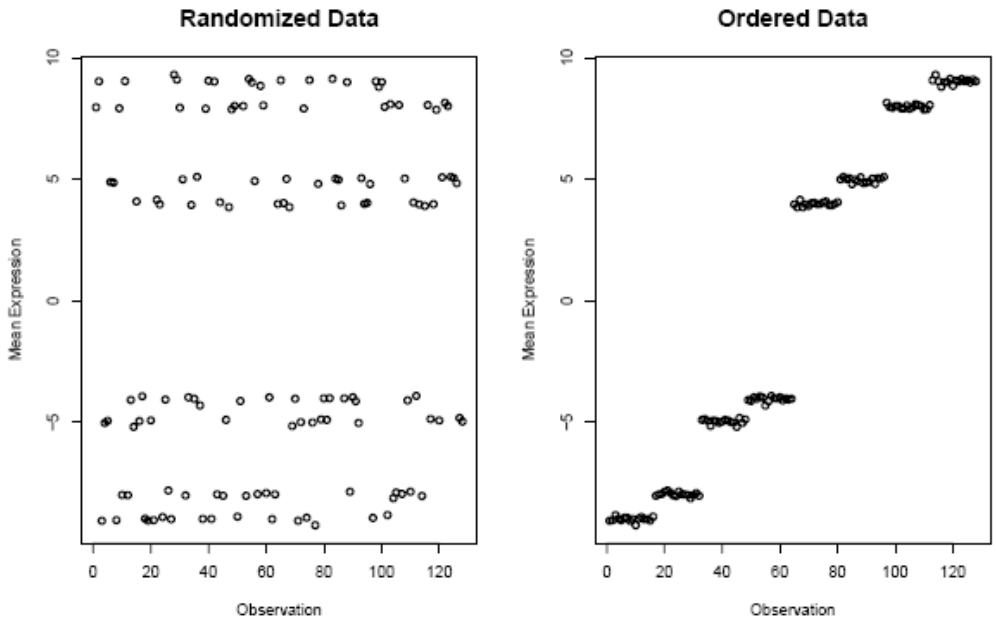


Figure 6 HOPACH ordering example, illustrating the difference between global and finer clustering structure.

3.2.2.3 **SPIN** (Sorting Points into Neighborhoods)

Compared to clustering algorithms, where the aim is to partition the data into several distinct groups, sorting points into neighborhoods is an iterative process to find an informative permutation of the data points. To find a meaningful order (needle in a haystack problem), that reveals the nature of an inherently continuous phenomenon (e. g. evolution of a certain disease) two different iterative search algorithms, Side to Side (STS) and Neighborhood, with O (n^2) and O (n^3) step-complexity are applied.

The input for SPIN is a symmetric distance matrix D of dimension n x n. The necessarily different permutations of D for the iterative process are created with the help of permutation matrix P. The quality of the ordering achieved with the algorithms is quantified through a cost function $F(P) = tr(PDP^TW)$, where *tr* indicates the trace of the resulting matrix and W is a weight matrix. In the case of STS algorithm the weight matrix becomes $W = XX^T$, with X being an increase in distance between loci (in our study with $X_i = i - (n + 1)$). For the neighborhood algorithm the

weight matrix is $W_{ij} = e^{-(i-j)^2/n\sigma}$, where σ is an indicator of the size of the neighborhood for which the ordering is optimized. The final results of SPIN are presented in a full pair wise distance matrix of the data points viewed in pseudocolor.

3.2.2.3.1 Side to Side (STS) Algorithm

Side to Side (STS) permutation constructs a grouping of loci by placing markers with high dissimilarity values far apart from those with high similarity values. With this process it is guaranteed that markers which are placed far apart in the linear ordering are also distant in full high dimensional space (Figure 7b). Thus, a permutation result achieved by the STS algorithm is similar to the projection on a principle components analysis (PCA) plot. PCA technique is a linear dimensionality reduction technique that seeks to identify a small number of components that capture most of the relevant structure in the data. With this linear dimensionality the loci can be distinguished into different groups from high similarity (blue colored) to low similarity (red colored) (Figure 7a). This PCA technique, applied in the SPIN software, has been already used to order a large number of genetic markers (e. g., thousands of SNPs) and to infer information between populations (Paschou *et al.* 2007).

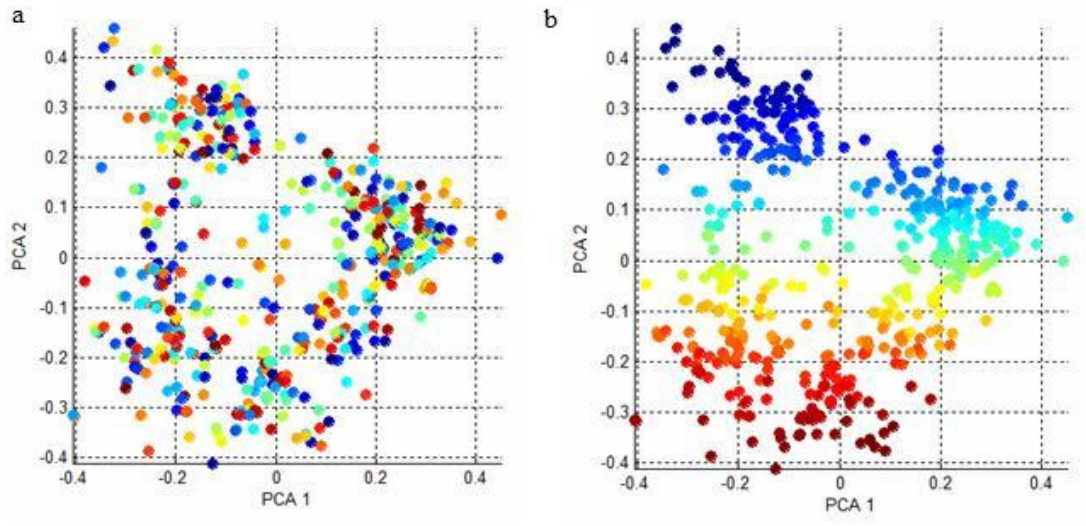


Figure 7 Projection of SNPs from BTA1 on the first and second PCA, (a) randomized data set of SNPs, (b) Order result after STS algorithm application.

Concerning the loci ordering problem (TSP) this algorithm is used to minimize the non deterministic polynomial (NP)- time complete problem of proving that a graph

contains different groups of elements (Garey & Johnson 1979). Hence, this algorithm is applied in advance of the Neighborhood algorithm to achieve a first partitioning of the data points, in this study SNPs (Assif Yitzhaky, personal communication).

3.2.2.3.2 Neighborhood algorithm

The minimized NP problem has been solved with the Neighborhood algorithm. This algorithm tries to improve the positioning of grouped loci by ensuring that loci with high LD (blue colored) are positioned near the main diagonal in the distance matrix. For orderings to be most accurate this algorithm has to start with σ=n (where n = data points of the distance matrix and σ= width parameter) and applied iteratively with stepwise reduction of the width parameter (in steps of 50 until 50, followed by values of 20, 10, 5 and 1). Figure 8 shows the result of the applied algorithms starting with a randomized distance matrix up to a well ordered distance matrix.

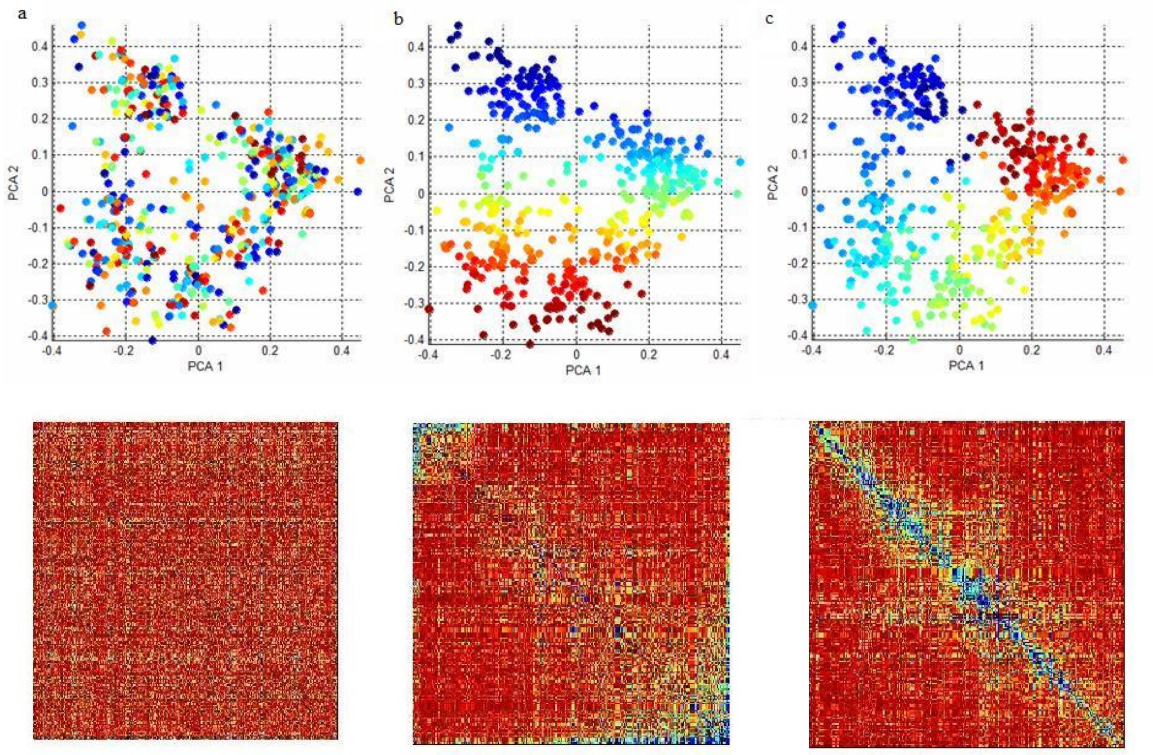


Figure 8 The different stages of an ordering process using the method SPIN, for BTA1, illustrated on the PCA scatter plot and corresponding distance matrix. (a) Randomized set of SNPs, (b) ordering result after STS application (c) Final ordered distance matrix with Neighborhood algorithm.

3.2.3 Data subsets

Initial investigations of the different order algorithms applied in this study showed that only SPIN was successfully adopted to compute the most likely order of a set of SNPs on BTA6 (Khatkar *et al.* 2006). The different method tests of SPIN on single bovine chromosomes (BTA1, BTA2, BTA4, BTA6) revealed that it is best to use both SPIN algorithms (STS and Neighborhood) on a distance matrix with D' values [2] to solve the TSP concerning the loci ordering. To infer the general features of the current LODE map approach presented in this study the SPIN algorithms were further applied on different sets of SNPs including whole genome- and minimum sample size analyses.

Concerning the allocation of unmapped SNPs to an order of SNPs with known positions, initial investigations of BTA1 showed that r^2 distances were successfully applied instead of D' distances to position unmapped SNPs on the current bovine map, as already shown by Miller and Hayes (2006). In this case only the neighborhood algorithm was applied to determine the most likely positions of the SNPs.

To test the ability of our current LODE map approach using genetic distances to position and order SNPs on the current bovine map (Btau3.1), the SPIN algorithms were applied to numerous test batches where "unknown" SNP positions had been set up. Five main objectives were considered regarding these test batches:

1. Is there a possibility to order SNPs in a whole genome using solely estimates of LD? (test batch 1)
2. Is there a main difference between Millers approach using r^2 values and the current LODE map approach using r^2 distances to determine the most likely SNP position? (test batch 2)
3. What happens with SNPs only genotyped in a few animals? (test batch 3)
4. Is there a common possibility, that SNPs with low MAF can not be aligned? (test batch 3)
5. Is it possible to align additional SNPs, which have not been aligned with other mapping strategy? (test batch 4)

3.2.3.1 Test batch 1 – Full bovine LODE Map

To investigate if the LODE map procedure can be applied on a full bovine genome a selected panel of 8,849 SNPs genotyped in the 1546 Holstein-Friesian bulls were included. For this initial study SNPs with minor allele frequencies (MAF) <0.05, as well as SNPs deviating form Hardy- Weinberg equilibrium (HWE) at a significance level of p<0.0001 were excluded. As a third criterion only SNPs at least genotyped in 1,240 bulls were used. To test the LODE map as a superior tool for positioning "Problem SNPs" a panel of 96 SNPs defined as being problematic in the current Btau3.1 assembly were distributed on all chromosomes, where the number per chromosome was ranging from 150 to 504. The mean marker interval on the chromosomes was within a range from 59 to 275 kb.

3.2.3.1.1 Fisher r-to-z transformation

Fisher r-to-z transformation was used for comparisons between the single mapping strategies (Btau3.1, CRC and LODE). This method calculates a value of z that can be applied to assess the impact of the difference between two correlation coefficients, r_a (Btau3.1 – LODE) and r_b (CRC – LODE). If r_a is greater than r_b, the resulting value of z will have a positive sign; if r_a is smaller than r_b, the sign of z will be negative. (http://faculty.vassar.edu/lowry/rdiff.html)

3.2.3.2 Test batch 2 – Alignment of high quality known SNPs as unknown

To test the ability of the SPIN procedure to position a set of n SNPs, a data subset of 270 known SNPs compromising nine approximately equally-spaced loci on each bovine chromosome was prepared and treated as "un-aligned with unknown positions". To investigate potential ordering problems for SNP with low MAF, we relaxed the criteria to include SNP with MAF>0.01. Next, we created a new set of LODE maps excluding these 270 loci. The set of 270 SNPs was then randomly permuted to simulate a set of 270 unaligned SNPs. For the proportion of known SNPs (chromosomes), the same panel of SNPs was used as in the full bovine LODE map above, which makes a total of 9,119 SNPs for this test batch.

3.2.3.3 Test batch 3 – Alignment of low quality and problem SNPs

To test the robustness of the alignment and ordering procedures, a test batch of SNPs with low significant LD due to small sample size and low extent of LD, was prepared. All SNPs with MAF<0.05 and SNPs only genotyped in a few animals were included. Furthermore, the CRC data set defined SNPs as being problematic if not in Hardy Weinberg equilibrium (HWE) or showing segregation distortion; these SNPs were also included in the test batch. In addition, a total of 428 known SNPs as for test batch 2 were included as internal controls. In total a set of 11,426 SNPs were included in this test batch.

3.2.3.4 Test batch 4 – Alignment of current unaligned SNPs

The fourth test panel of SNP was selected to check on the robustness of the CRC consensus integrated mapping approach, and to align SNPs which were currently unaligned by the integrated consensus mapping approach. In this instance, a test panel of 640 SNPs comprised 400 SNPs with known CRC positions and 240 current unaligned SNPs. The CRC quality table concerning the 640 SNPs classified 21 SNPs as "Conflict SNPs" due to the different positions of the genetic maps used for the consensus mapping strategy, and 60 SNPs with a MAF<0.05. SNPs genotyped in a few animals were not included in this test batch.

3.2.4 The calculation of SNP position

To identify the physical positions of the aligned SNP's, the LD distances between companion neighbors were used to position SNP i as

$$pos_i = pos_{i-1} + \frac{d_1}{d_1 + d_2}\left(pos_{i+1} - pos_{i-1}\right) \quad (3)$$

With pos_{i-1} and pos_{i+1} being Btau3.1 positions of the neighboring SNP and d_1 and d_2 the corresponding LD distances.

4. RESULTS

4.1. Ordering algorithm comparison

To investigate the best ordering algorithm for solving the TSP, initial orders with D' distances of BTA6 were calculated and compared to the Btau3.1 map order. The orderings of the single algorithms show, that only SPIN and Greedy algorithm, improve ordering (part of HOPACH), can be applied to generate meaningful orders of the SNPs (Figure 9). However, the Greedy algorithm was only run with a sub set of 20 SNPs, due to capacity limitations of this process.

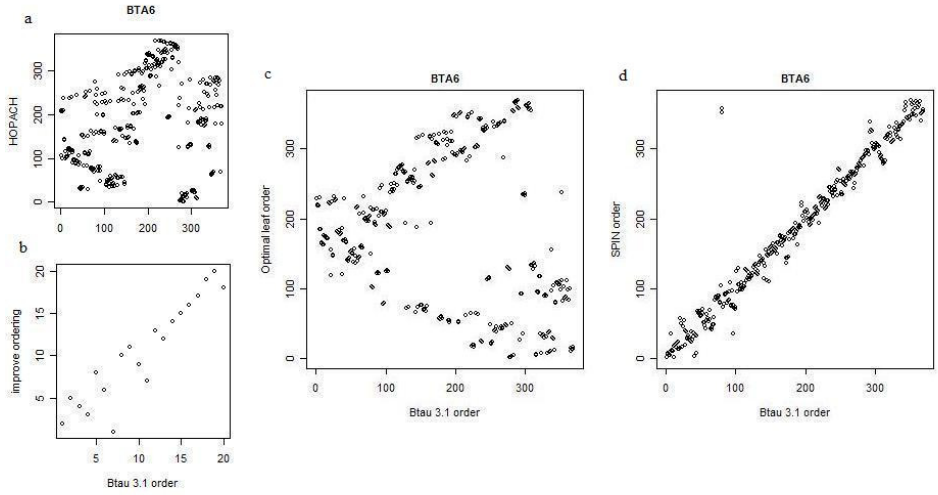


Figure 9 Grids comparing the ordering results of the different applied algorithms with Btau3.1 locus order for BTA6, (a) using HOPACH, (b) using step-wise algorithm (`imroveordering`), (c) using Optimal leaf order and (d) using SPIN method.

4.2. Method Tests

A resample technique based on different random starting points of loci was chosen to test the LODE map approach using SPIN algorithms for robust repeatable outcomes. The results from different runs have shown that using D' is clearly better than using r^2 in calculating marker order. For the different permutations slight variations in using D' distances and swapped final orderings have been noted. The low correlation using r^2 distances is most likely due to an incorrect placement of SNPs at both ends of the LODE order.

Table 1 Rank correlations of Btau3.1 and LODE order of 5 different runs, using D' distance and r^2 distances form BTA1, 2 and 6.

BTA1 (452 SNPs)		BTA2 (412 SNPs)		BTA6 (369 SNPs)	
D'	**r^2**	**D'**	**r^2**	**D'**	**r^2**
0.9577504	0.9184613	0.9798238	-0.2715495	-0.9670135	0.7264980
-0.9576643	-0.3534507	-0.9796840	-0.3274285	0.9670430	0.7741865
-0.9576644	-0.3035515	-0.9796840	-0.5482516	0.9705218	0.8407001
0.9575506	-0.7960630	-0.9796843	0.7574576	0.9705218	-0.7273299
-0.9576644	-0.3036974	0.9798238	0.8168717	-0.9670135	0.8166550

The SPIN application provides the ability to play around with the algorithms (STS and Neighborhood). To minimize the need of the algorithms to get an optimized method for ordering SNPs an additional method only using the Neighborhood algorithm was tested. With the use of this method meaningful orders were obtained, but the results were not as convincing, as those obtained using both algorithms (STS & Neighborhood), even after multiple runs (Figure 10).

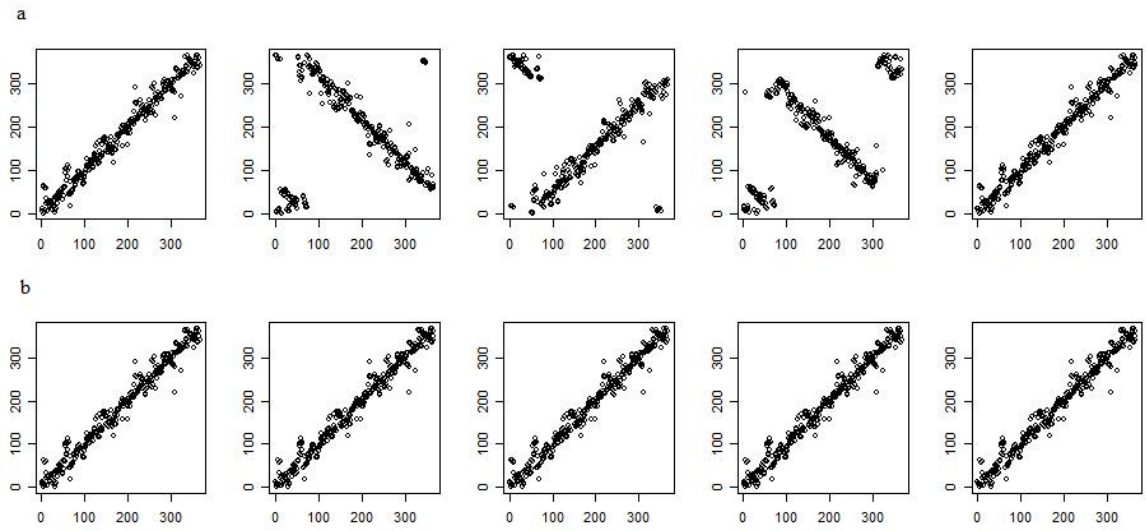


Figure 10 Grids comparing BTA4 LODE map with Btau3.1 locus order after 10 different starting points. (a) using only SPIN algorithm Neighbourhood, (b) using both algorithms (Side to Side and Neighbourhood).

4.3. Minimum sample size

To determine the minimum sample size required to achieve a reasonable ordering, pair-wise LD distances for random subsets of 1,000, 500, 400, 300, 200, 150, 100, 75, 50 and 25 of the 1,546 bulls were calculated, applying the same limits with respect in MAF and HWE as in the test chromosomes. The ordering was consistently good using 500 individuals, reasonable using 150 and very poor for sample sizes below 100. A typical result is illustrated in Figure 11a. The effect of marker spacing was checked by randomly sampling 300, 150, 75 and 25 SNPs from a single chromosome. The results for BTA2 (Figure 11b) indicate that the procedure works very well when considering 75 SNP per chromosome. The split into two separate blocks in the case of 150 SNPs did not occur in several other random samples.

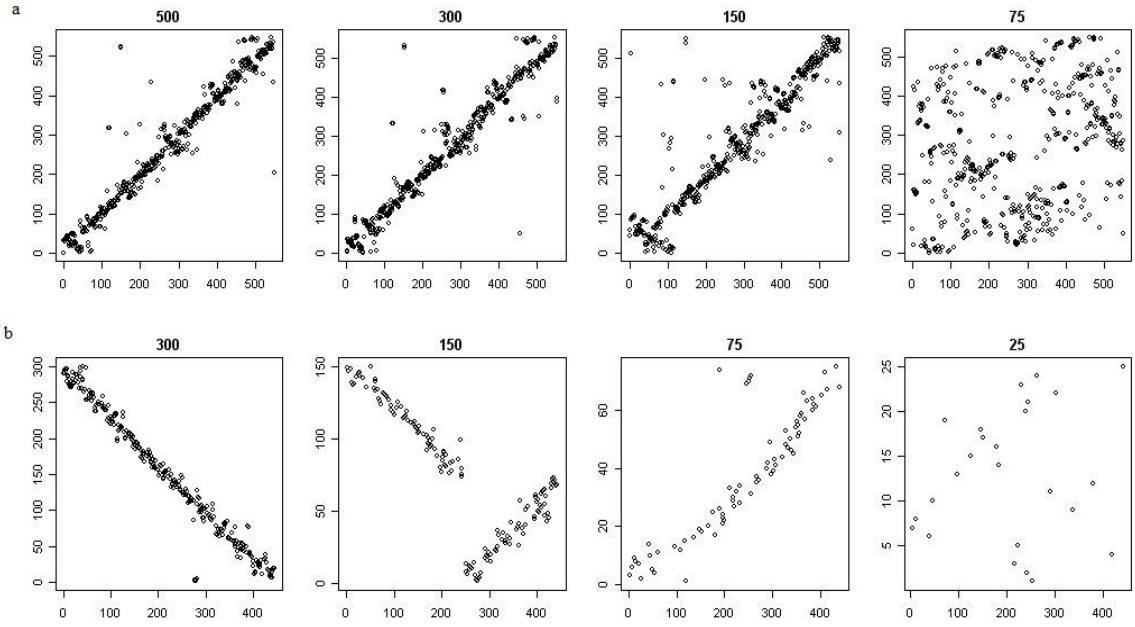


Figure 11 Effect of sample size and number of loci per chromosome. (a), Grids comparing locus order in bovine LODE maps of BTA1 with Btau3.1 locus order for 500, 300, 150 and 75. (b) Grids comparing locus order in bovine LODE maps of BTA4 with Btau3.1 locus order for 300, 150, 75 and 25 randomly selected SNPs per chromosome.

4.4. Whole genome analyses

To extend the ordering strategy to a whole genome analyses, pair-wise D' values were calculated for all SNPs on five autosomes (the two largest, one intermediate, and the two smallest chromosomes: BTA1; BTA2; BTA14; BTA28 and BTA29 respectively). As shown in Figure 13 the SPIN algorithm did an almost perfect job of allocating loci to chromosomes, and succeeded in creating a very accurate ordering for each chromosome studied. To distinguish the different autosomes without knowing the Btau3.1 order, the Score and diagnostic plot (part of the SPIN software) were used. However, at this plot 6 groups are distinguished, which posed a problem (Figure 12b). This was due to the separation of BTA1 in two blocks of SNPs. This break of BTA1 has also been noted in single runs, due to the window size of Figure 12a it is not visible in a panel of five autosomes.

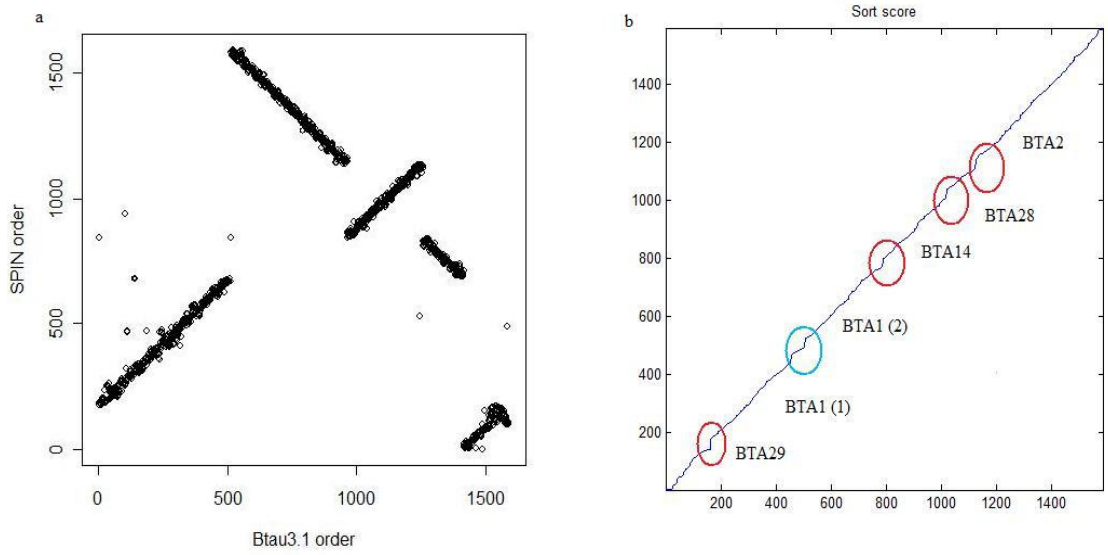


Figure 12 Result of the separation for a pool of SNP markers from five chromosomes. (a) Grid of SPIN order versus Btau3.1 locus order, (b) Diagnostic and Sort score plot of the order result.

4.5. Alignment procedure

To investigate that besides r^2 values also D' values could be possible used to align current unaligned SNPs. The neighborhood algorithm was applied on a distance matrix of BTA1 using r^2 and D' distances as well. Therefore, the BTA1 SNP subset was assembled of a proportion of 504 SNPs known form the Btau3.1 map and 270 "unknown SNPs" compromising 9 SNPs associated to the BTA1 chromosome. This initial alignment result of BTA1 shows that with D' distances these SNPs could not been separated from the others. Hence, whole scaffolds of SNPs have been aligned to the chromosome (Figure 13a). In contrast to the alignment result using r^2 distances, where 8 out of 9 SNPs were aligned to the BTA1 chromosome (Figure 13b).

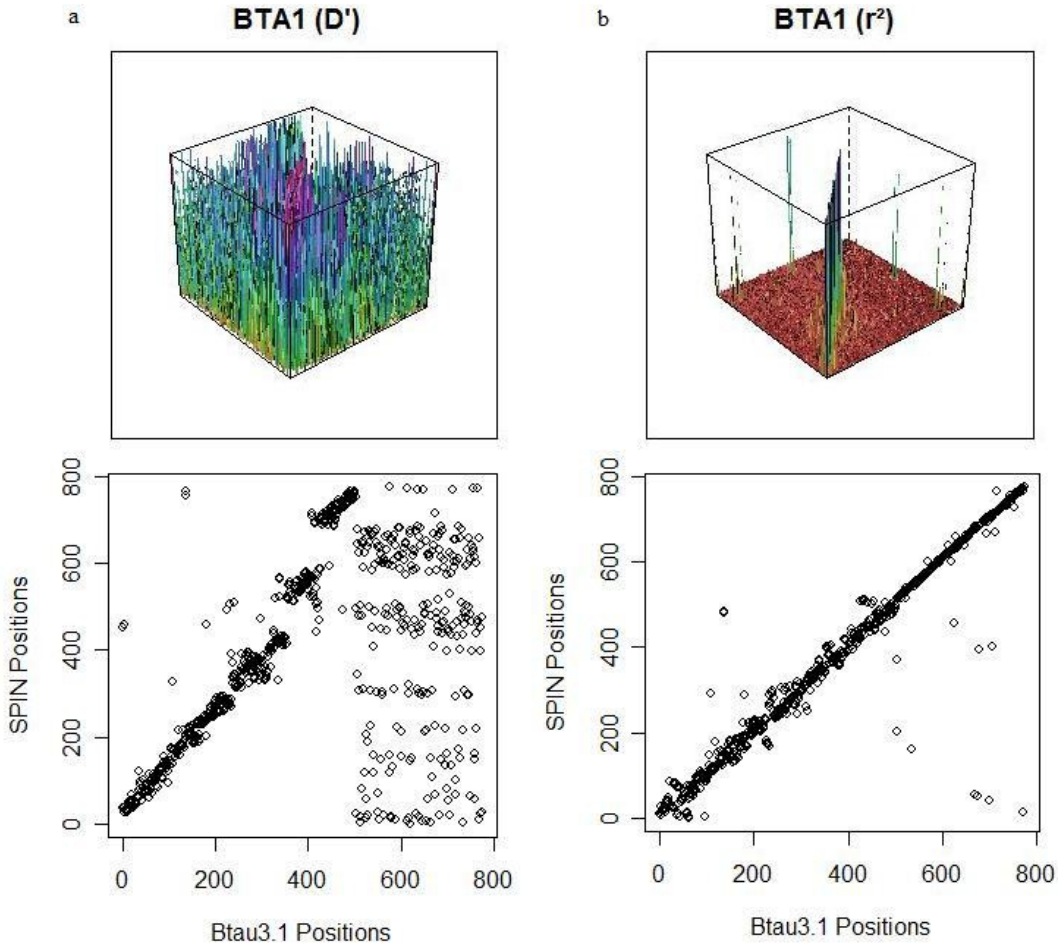


Figure 13 3D scatter plots indicating the extent of LD and position results of BTA1 adding 270 unknown SNPs. (a) Positioning result using D' distances. (b) Positioning result using r^2 distances.

4.6. Test batch 1 – Full bovine LODE Map

The LODE map order for each single bovine chromosome was determined after 10 permutations. These resultant orderings were compared with Btau3.1 assembly (Table 2). A significant variation between repeats was noted only at BTA30 ($p < 0.005$). An identical ordering was achieved at 12 chromosomes only. The absolute rank correlations of LODE with Btau3.1 orders ranged from 0.827 to 0.995 for 29 chromosomes, with an average of 0.947, weighted by the number of SNPs per chromosome, excluding BTA10.

Table 2 Absolute rank correlations between LODE map orders and Btau3.1 orders.

Chromosome	SNPs	SNP interval	Btau3.1 LODE	Chromosome	SNPs	SNP interval	Btau3.1 LODE
		(kb)				(kb)	
BTA1	504	103	0.970	BTA16	299	87	0.963
BTA2	444	104	0.995	BTA17	284	85	0.970
BTA3	452	107	0.935	BTA18	273	79	0.965
BTA4	367	130	0.986	BTA19	326	65	0.980
BTA5	393	113	0.904	BTA20	245	85	0.879
BTA6	421	122	0.938	BTA21	170	229	0.913
BTA7	342	131	0.991	BTA22	233	87	0.971
BTA8	372	116	0.964	BTA23	246	59	0.940
BTA9	251	193	0.958	BTA24	209	124	0.957
BTA10	379	118	0.240	BTA25	198	77	0.912
BTA11	427	116	0.975	BTA26	171	126	0.955
BTA12	265	119	0.895	BTA27	147	80	0.948
BTA13	388	70	0.965	BTA28	150	87	0.964
BTA14	285	79	0.975	BTA29	165	107	0.910
BTA15	293	103	0.827	BTA30	150	275	0.901

In case of BTA10 no correlation was observed due to two single linked scaffolds (blue circled in the correlation plot). The problem concerning these groups of SNPs is that they do not have any association to SNPs in the neighborhood (Figure 14a). Through this pattern of LD the Side to Side algorithm is providing a wrong grouping result, by placing these SNPs to SNPs located on the end of the chromosome. Hence,

a subset of BTA10 has been created excluding these groups of SNPs. Using this subset of BTA10 a correlation of 0.966 for this chromosome was achieved (Figure 14b), compared to 0.240 for Btau3.1.

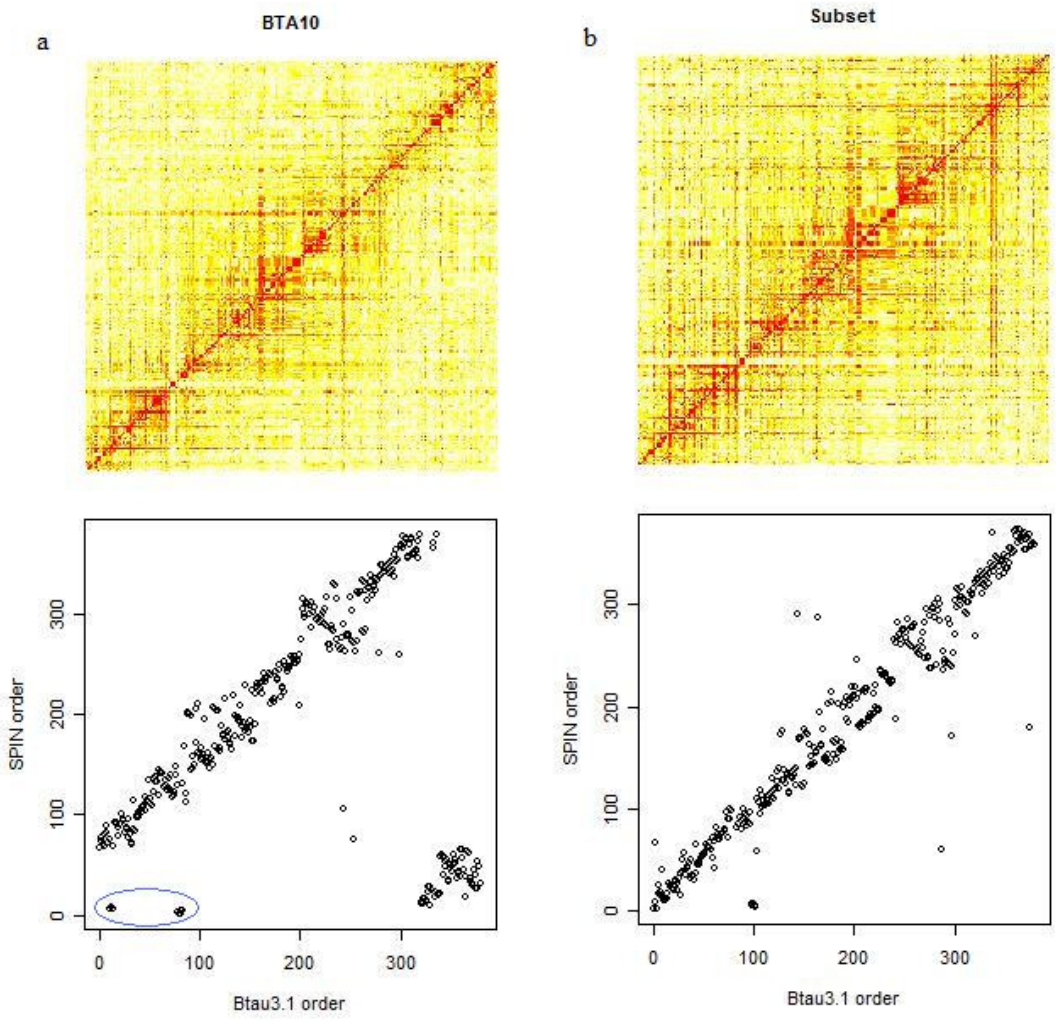


Figure 14 Heat maps and ordering results for BTA10, (a) including separated linked scaffolds, (b) using a subset of SNPs (excluding these scaffolds).

An oxford grid between the LODE map orders and Btau3.1 orders illustrates, that no SNP (orphan) has been placed in conflict with the Btau3.1 position along the whole bovine genome (Figure 15). Within the single chromosomes about 100 LODE positions significantly deviated from those obtained through Btau3.1 assembly. The prior assigned 96 “Problem SNPs” across the entire genome have been positioned in close proximity to SNPs in the neighborhood. Hence, these SNPs have not caused any errors in the LODE map approach.

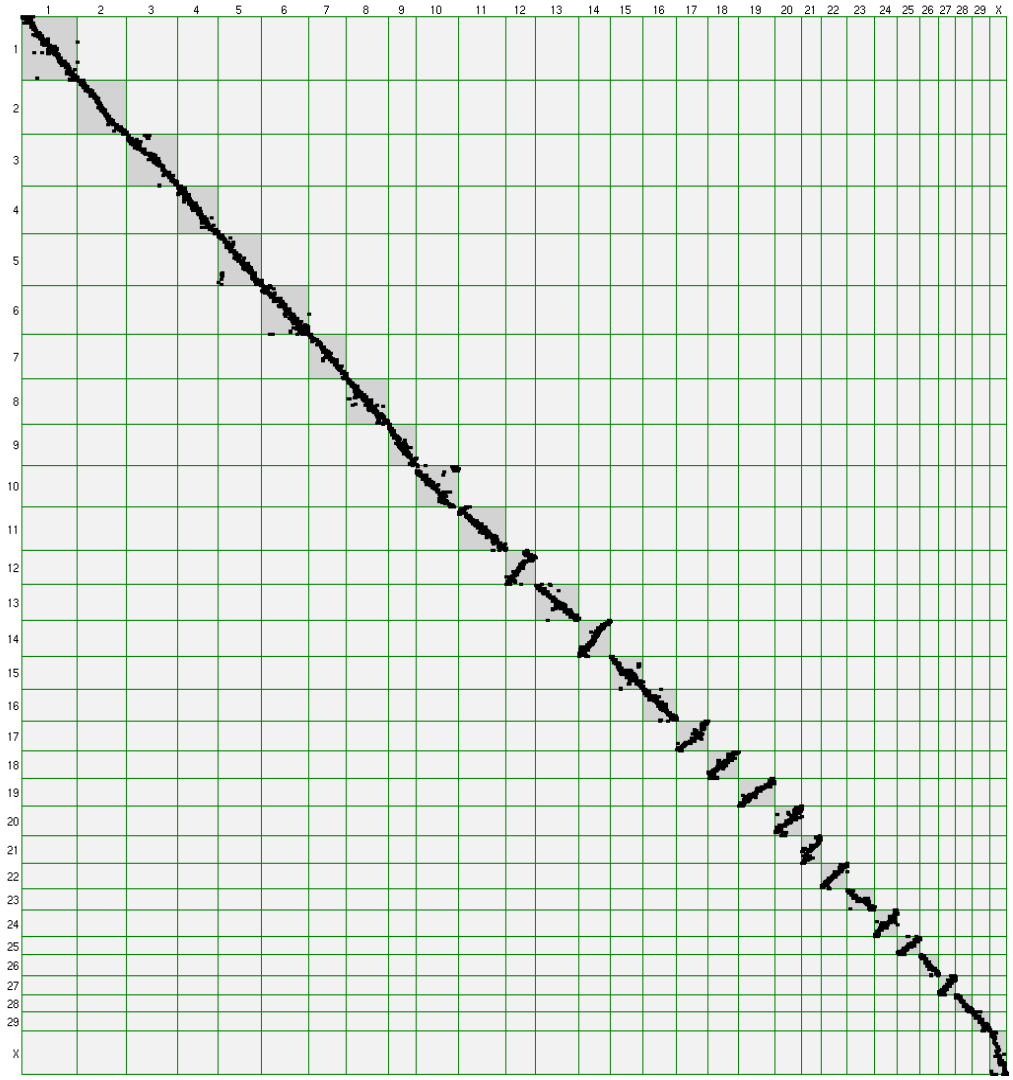


Figure 15 Oxford grid presenting the absolute rank correlations between the LODE map orders and Btau3.1 map orders.

4.6.1 Three map comparison

At the time of analysis the Btau4.1 assembly was not available, so the results were compared with the CRC assembly, which shows considerably better agreement with individual bovine maps (see Chapter 3). In this map comparison, the absolute rank correlations between the three maps (Btau3.1, CRC and LODE) were calculated. To evaluate the agreements between the different maps Fisher-r-to-z transformation was used.

Table 3 Absolute rank correlations between Btau3.1 orders, CRC orders and LODE orders.

Chromosome	SNPs	SNP mean interval (kb)	Btau3.1-CRC	Btau3.1-LODE	CRC - LODE	z	p
				r_a	r_b		
BTA1	504	103	0.987	0.970	0.982	-3.96	0.0001
BTA2	444	104	0.998	0.995	0.995	-0.49	0.6241
BTA3	452	107	0.997	0.935	0.932	0.37	0.7114
BTA4	367	130	0.992	0.986	0.987	-0,62	0.5353
BTA5	393	113	0.999	0.904	0.905	-0.06	0.9522
BTA6	421	122	0.989	0.938	0.948	-1.31	0.1902
BTA7	342	131	0.999	0.991	0.991	-0.31	0.7566
BTA8	372	116	0.991	0.964	0.973	-2.05	0.0404
BTA9	251	193	0.996	0.958	0.960	-0.37	0.7114
BTA10	379	118	0.998	0.240	0.241	-0.19	0.8493
BTA11	427	116	0.998	0.975	0.979	-1.20	0.2301
BTA12	265	119	0.995	0.895	0.896	-0.13	0.8966
BTA13	388	70	0.999	0.965	0.965	-0.07	0.9442
BTA14	285	79	0.998	0.975	0.975	-0.10	0.9203
BTA15	293	103	0.881	0.827	0.959	-9.09	0.0000
BTA16	299	87	0.998	0.963	0.964	-0.16	0.8729
BTA17	284	85	0.997	0.970	0.972	-0.39	0.6965
BTA18	273	79	0.988	0.965	0.970	-0.99	0.3222
BTA19	326	65	0.999	0.980	0.981	-0.30	0.7642
BTA20	245	85	0.974	0.879	0.908	-1.56	0.1188
BTA21	170	229	0.930	0.913	0.941	-1.86	0.0629
BTA22	233	87	0.998	0.971	0.980	-1.89	0.0588
BTA23	246	59	0.990	0.940	0.944	-0.36	0.7188
BTA24	209	124	0.997	0.957	0.960	-0.37	0.7114
BTA25	198	77	0.989	0.912	0.922	-0.59	0.5552
BTA26	171	126	0.998	0.955	0.956	-0.14	0.8887
BTA27	147	80	0.999	0.945	0.944	0.06	0.9522
BTA28	150	87	0.997	0.964	0.967	-0.34	0.7339
BTA29	165	107	0.992	0.910	0.906	0.18	0.8572
BTA30	150	275	0.965	0.901	0.921	-0.78	0.4354

The negative z value in 27 of 30 cases indicates that generally LODE map orders better agree with CRC orders (r_b) than with Btau3.1 orders (r_a). A significantly better ordering can be achieved at the BTA1 and BTA15 ($p < 0.0001$). Only at three

chromosomes, BTA3, BTA27 and BTA29, the correlation between orders did decrease compared to the Btau3.1 order, but this was not significant ($p < 0.71$). In addition through the use of CRC position as evidence the numbers of SNP positions in conflict to Btau3.1 positions (orphans) within chromosomes have been reduced to 80 SNPs.

When the LODE maps were compared with the CRC assembly, the average rank correlations, weighted by the number of SNPs per chromosome, is 0.956 (ranging from 0.864 to 0.995 with the exception of BTA10). This result is a significant ($p<0.0001$) better than the correlation of the LODE maps with Btau3.1 (Figure 16).

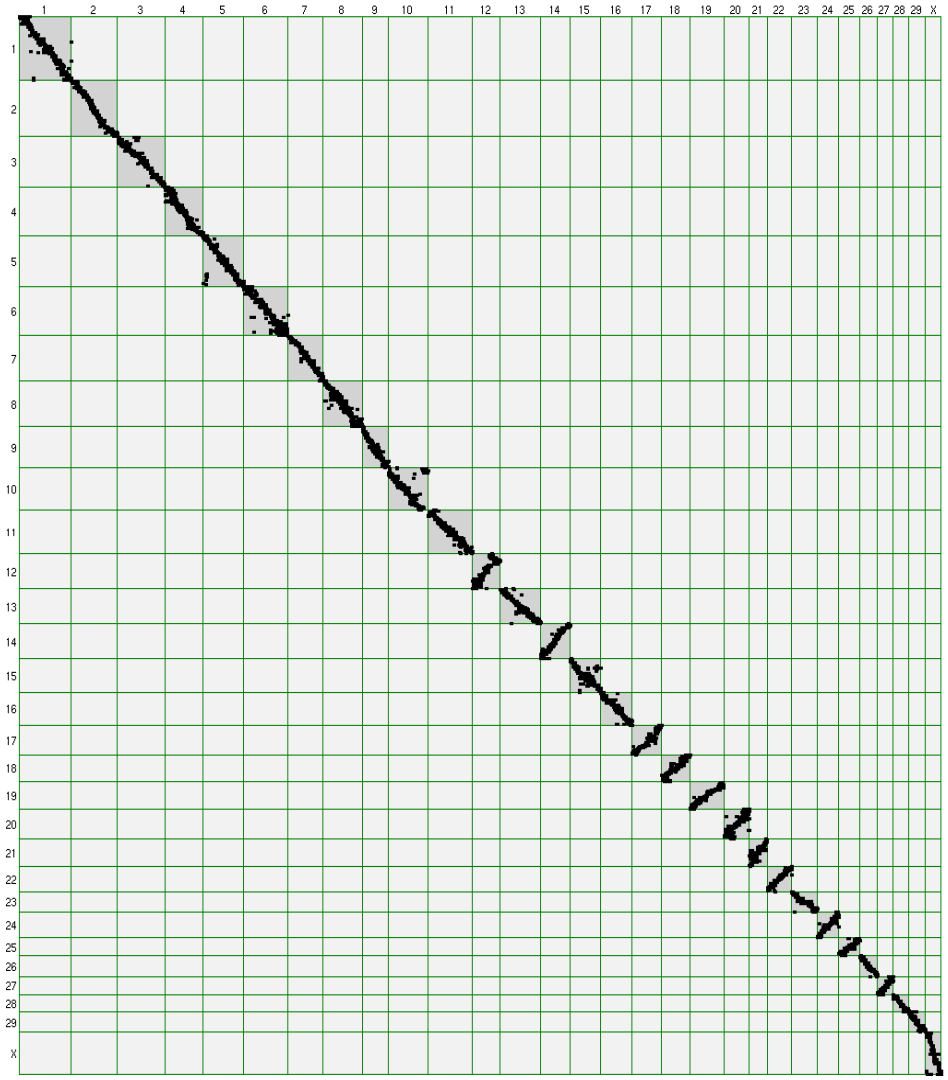


Figure 16 Oxford grid presenting the absolute rank correlations between the LODE map orders and CRC orders.

4.7. Test batch 2 – Alignment of high quality known SNPs as unknown

The LODE procedure concerning the alignment (neighborhood algorithm and r^2 distance matrixes) has proven efficient in the positioning of SNPs. Of the 218 SNPs placed on 30 chromosomes, 216 (80%) of them were on the correct chromosome, according to Btau3.1. One SNP was placed on the wrong chromosome (BTA12 instead of BTA8) and another one was placed on two chromosomes (BTA13 and BTA11, where BTA11 is the true one) compared to the Btau3.1 alignment. Of the 52 non-placed SNPs, 19 showed a MAF<0.05, of which 2 SNPs were positioned on the true chromosome. With restriction on quality, the result of truly positioned SNPs increased to 86% (213 out of 248 SNPs).

A second subset of SNPs, where the orderings of chromosomes have been created with the SPIN algorithms, was chosen to test the quality of the alignment procedure by reducing prior physical information. In this case the success of the numbers of correctly aligned SNPs was slightly decreasing. Of the 216 SNPs placed on the chromosomes, 213 (78%) were aligned to the correct one. Only 3 SNPs have been aligned on two chromosomes. No SNP was placed on the wrong chromosome.

The median distance of the approximated positions from the Btau3.1 positions of the 218 SNPs was 312 Kb, with 5% and 95% quantiles of 0.18 Kb and 5.91 Mb. This procedure was superior to approximating the position by averaging the positions of the two nearest neighbors (median distance of 1.09 Mb), than averaging the positions of the 4 or 6 nearest neighbors according to their SPIN order, considering a larger neighborhood when placing unaligned SNPs (median distances of 1.48 and 1.70 Mb). Treating the r^2 distances as squared distances and replacing d1 and d2 by their square roots did not improve the positioning (median distance of 470 Kb). Concerning the second subset the median distance of the 216 approximated positions was slightly increased by using LODE orders instead of Btau3.1 orders. In this case the correlations between the Btau3.1 and approximated positions decreased from 0.996 to 0.995.

4.8. Test batch 3 – Alignment of low quality and problem SNPs

This study demonstrates that including low quality SNPs (only genotyped in few animals) or SNPs already indicated as "Problem SNP" by scientists in the 10k parallel panel caused a lot of errors in the alignment process. Especially these SNPs were aligned on multiple chromosomes (Table 4) and incorrectly placed compared to current Btau3.1 map positions (Table 5).

Table 4 SNPs assigned on multiple chromosomes

Assay-ID	SNP aligned on Chromosomes	nAA	nAB	nBB	Numbers of bulls	MAF	Problem SNP
342632	11,3	1254	274	13	1541	0.097	
342713	11,19	44	432	1064	1540	0.169	
344717	19,26	700	706	137	1543	0.312	
351039	26,5	123	676	743	1542	0.078	
342585	1,11,27	673	709	158	1540	0.333	
347562	6,7,19,26	1145	372	11	1528	0.123	Yes
343977	2,3,7,11,15,24	12	48	47	107	0.336	Yes
342512	11,15,19,23,26	918	617	3	1538	0.203	Yes
347526	1,6,7,11,19,26	0	100	1	101	0.495	
350851	6,7,11,18,19,24,26	7	89	0	96	0.464	
343348	2,3,5,6,7,15,18,23,26	3	61	28	92	0.364	Yes

Among the incorrectly placed SNPs, 6 had a good quality score. To control if they also would be placed to the correct physical position without low quality SNPs in the neighborhood another subset for these 6 SNPs was prepared. The alignment result of these SNPs showed, that two (35351 and 348913) were re-aligned on the same Btau3.1 chromosome, only one (349931) was placed on a different chromosome (BTA5 instead of BTA10) and the others (353276, 350927 and 343959) are no longer aligned.

Table 5 Incorrectly placed SNPs

Assay-ID	Btau 3.1 chromosome	nAA	nAB	nBB	Number of bulls	MAF	Problem SNP
349162	2	5	81	0	86	0,471	
348307	8	11	81	0	92	0,440	
346249	14	4	90	1	95	0,484	
345111	5	2	96	0	98	0,490	
462300	13	1	52	45	98	0,276	
464305	8	1	98	0	99	0,494	
346730	18	0	98	1	99	0,495	
462298	5	0	96	4	100	0,480	
351244	25	11	90	0	101	0,445	
466288	10	4	98	0	102	0,480	
352606	5	0	89	14	103	0,432	
465897	5	35	69	0	104	0,332	
343149	20	31	75	0	106	0,354	Yes
346214	5	1	101	5	107	0,481	
466057	20	0	100	7	107	0,467	
353276	14	1025	192	127	1344	0,166	
350927	13	16	302	1213	1531	0,109	
348913	8	996	488	47	1531	0,190	
343959	5	339	777	416	1532	0,474	
353051	4	435	740	358	1533	0,475	
349931	10	479	1064	0	1543	0,345	

Despite the use of poor quality SNPs, 258 out of 428 SNPs (61 %) have been placed on one of the 30 chromosomes. To check the quality of the positions of this alignment, the result was compared with CRC integrated map positions, where 253 of the aligned SNPs have a high correlation (0.996) to CRC positions. From the SNPs with MAF>0.05 only 10 out of 41 SNPs have been assigned to single chromosomes.

4.9. Test batch 4 – Alignment of current unaligned SNPs

Out of the SNP subset with known positions (n=400), 301 SNPs were placed on the correct chromosome showing a high correlation (0.980) with CRC positions (Figure 17). Concerning the positions of 6 SNPs positioned by LODE gave a different result to CRC map positions and in 4 cases LODE aligned SNPs to two chromosomes. The difference (n=89) were not positioned by our LODE map procedure. Form the 240 SNPs without any prior positioning based on the integrated consensus map, 160 SNPs could be positioned with an r^2 range from 0.12 to 1. Only one SNP was placed on to two chromosomes. In this case it was also possible to align 12 SNPs, which were not assigned by CRC and thus marked as "Conflict SNP". Concerning the 60 SNPs with MAF<0.05 only 14 out of 60 SNPs have be positioned with the LODE map approach.

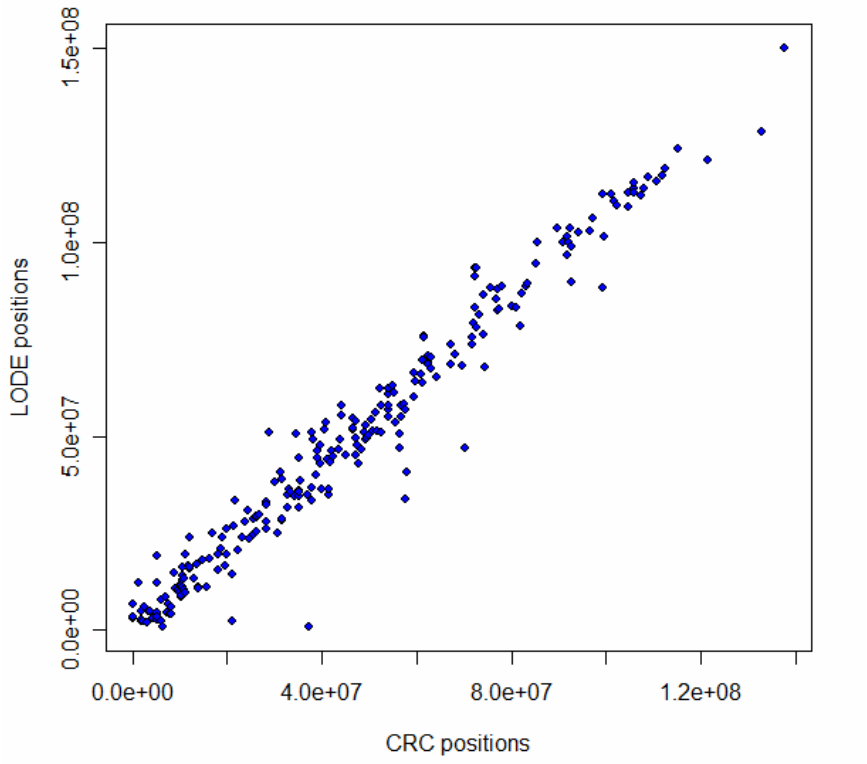


Figure 17 Correlation between 301 SNP positions mapped with the LODE Map strategy and CRC positions

5. DISCUSSION

5.1. GENERAL FEATURES of the LODE MAP

This study has shown that it is possible to order and position genetic markers (SNPs) based on LD with the aid of order algorithms. Creating a locus order of SNPs based on LD or genetic linkage is associated with the TSP problem. In this work, this NP (non-deterministic polynomial time) problem concerning the loci ordering has been solved successfully with a novel unsupervised approach SPIN (Tsafrir *et al.* 2005).

The results of the different method tests with SPIN have shown that both algorithms should be applied to receive robust repeatable outcomes. Hence the current method suggests that, it is best to do an initial partition of the data point (SNPs) in the distance matrix with the STS algorithm followed by the iterative ordering process using the Neighborhood algorithm. In this case only D' distances in the current cattle data set have the necessary properties to create meaningful orders. The initial ordering results have shown that with r^2 distances especially the endings of the LODE orders could not have been computed correctly, with the applied SPIN method. It was obvious that the nature of r^2 was responsible, that SNPs with high distances could not be separated from those with low distances (Humphreys 2007) (Figure 18b). A successful application of our LODE procedure can be predicted, if there is a clear separation of the SNPs in different groups, after the STS run such as in BTA1 (Figure 18a).

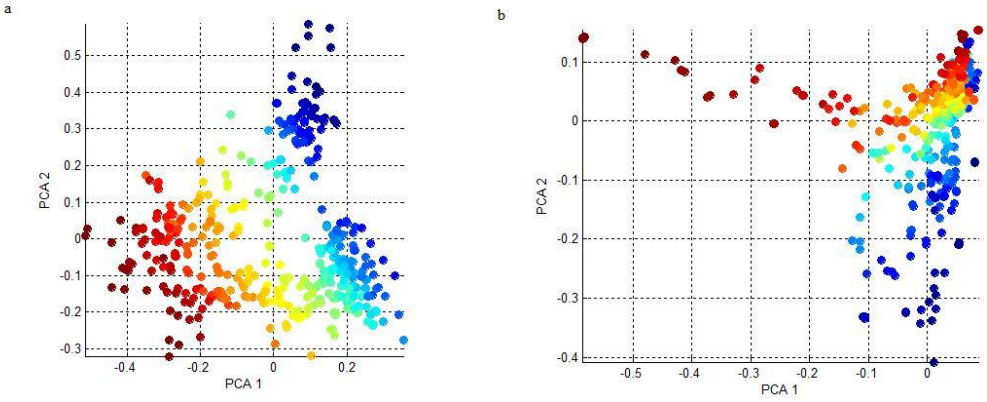


Figure 18 PCA scatter plots illustrating the order results after STS algorithm application, (a) using D' distances of BTA 1 in Australian Dairy Cattle, and (b) using r^2 distances of BTA 1 in Australian Dairy Cattle.

The investigations concerning minimum sample sizes of SNPs on BTA1 and the use of SPIN in a human population (Fung *et al.* 2006) show, that the general features of the LODE procedure are obviously dependent on the sample size and the extent of LD (Weiss & Clark 2002). In this case the PCA results of the STS runs of both samples already indicates, that an application of SPIN to order data sets described through low significant LD will not be effective (Figure 19).

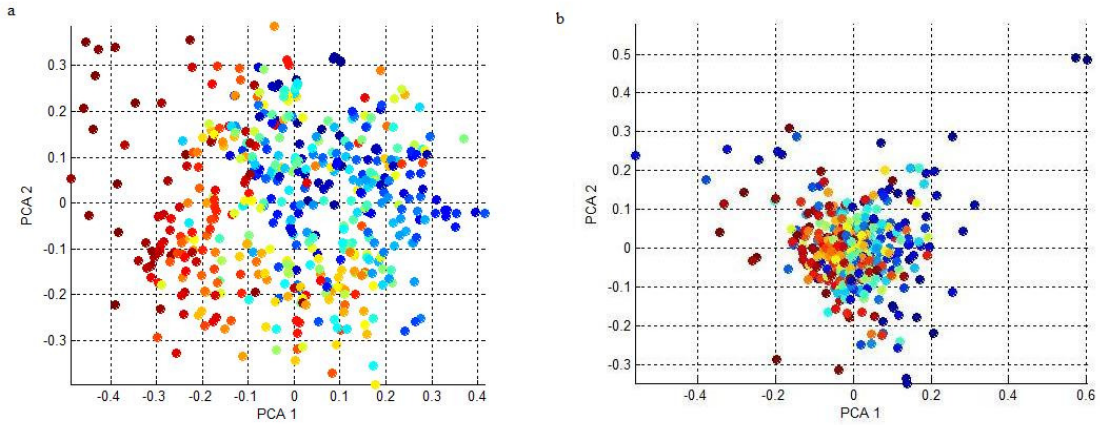


Figure 19 PCA plots illustrating the order results after STS application, (a) using a data set of SNPs of BTA1 only genotyped in 75 bulls, and (b) using a set of SNPs derived from HSA15 in the Parkinson disease population covering 100 Mb.

To investigate the results of the human population further, we prepared four different subsets of 500 SNPs genotyped in the Parkinson disease population on HSA15 (Fung *et al.* 2006). To increase the extent of LD we set different ranges for each subset, starting with 100 Mb down to 2.5 Mb. An application of the LODE procedure on the different test sets shows, that at a chromosomal region covering 5 Mb, the extent of LD in the human population becomes useful to create orders of SNPs in separated blocks (Figure 20). However, the extent and statistical significance of LD is too small to generate a high quality LODE order for this population.

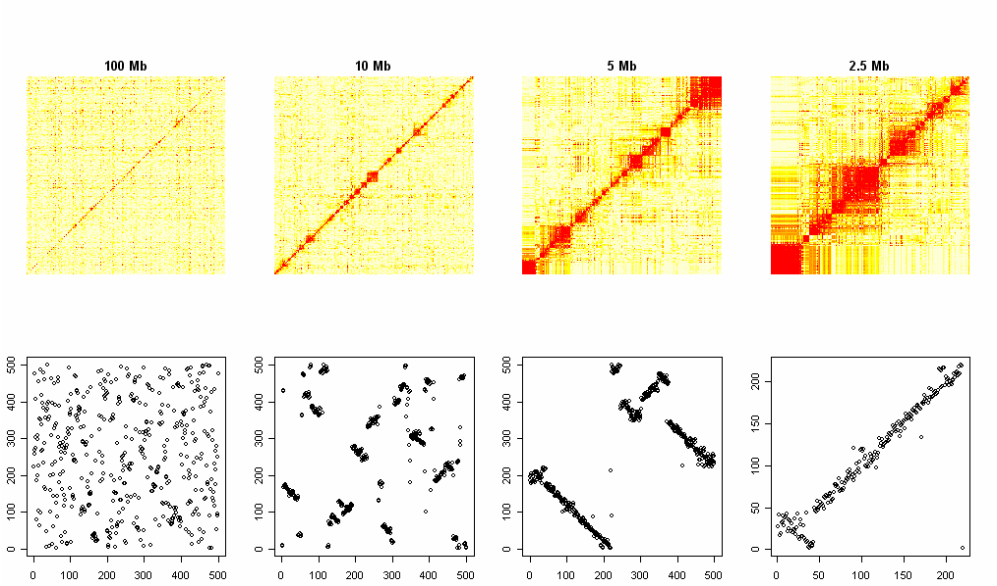


Figure 20 Heatmaps and LODE map results for the different SNP subsets on HSA 15 starting with 100 Mb down to 2.5 Mb

Hence the present results suggest that, for a useful implementation in whole genomes, LD needs to be of a similar extent to that in the Australian Dairy cattle population, where useful LD extends up to 18 Mb (Khatkar *et al.* 2006). To strengthen the statistical variance of LD, the SNPs should be genotyped in at least 1,000 individuals.

Considering the results of the alternative ordering algorithms (HOPACH and Fast optimal leaf order), it has been noted that through order algorithms based on clustering approaches, the TSP concerning LD could not have been solved successfully. In this case, the algorithm applied in the HOPACH package (namely the greedy, step-wise algorithm *"improveordering"*) could have been successfully used to create meaningful orders for a reduced subset of 20 SNPs. However, further investigations with other locus-ordering algorithms based on the solutions to the TSP (Faraut *et al.* 2007) need to be performed. Algorithms that have already been applied to linkage analysis like *seriation* (Buetow & Chakravarti 1987), *minimum sum of adjacent recombination fractions* (SARF) (Falk 1989) and *minimum product of adjacent recombination fractions* (PARF) (Wilson 1988) should also be tested. Such a comparison needs to be feasible for large numbers of loci. While SPIN was run within minutes on a moderately fast PC even for the largest problem presented (combination of the loci of 5 chromosomes), inclusion of all 15,036 bovine loci simultaneously would have been beyond the scope of the present version of the software.

5.2. SPECIFIC FEATURES of the LODE MAP

5.2.1 Order procedure

With the results of the first full bovine LODE map (test batch 1) in detail, the specific features of the current LODE map approach could have been further improved. In this case, especially the order results of BTA10 and BTA30 revealed new cognitions about the utility of a LODE map implementation.

The order result of BTA10 has shown that, the current ordering method becomes ineffective to generate a final order, if the LD consists of independent linked groups without any relations to other groups of SNPs in the neighborhood. An extreme example of this phenomenon is illustrated in a data set derived from 2,002 heterogeneous stock mice (Valdar *et al.* 2006). Generally, this mouse data set is described as expressing a high extent of LD assembled in independently linked LD groups. The high extent of LD in this population resulted in a higher accuracy of SNP positioning. However, the independent linkage groups suggest a final order in separated groups (Figure 21).

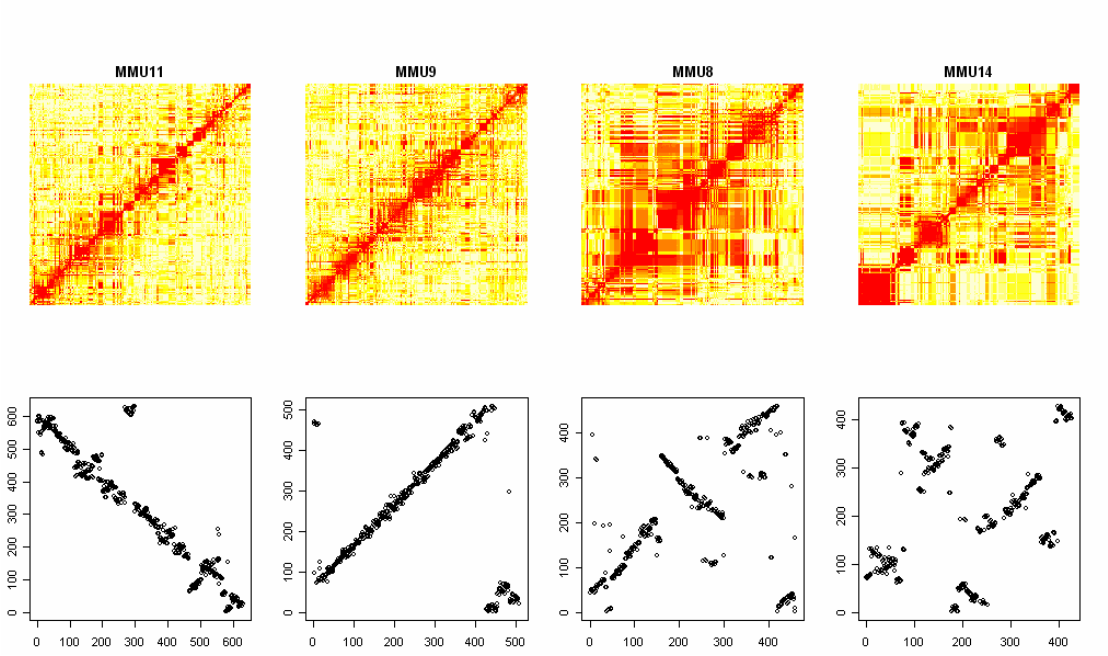


Figure 21 Heat maps (distance matrixes) and order results of the LODE map strategy from selected mouse chromosomes.

The method test with reduced SNP marker density for BTA2 shows that the reproducibility of the order results is obviously dependent on the numbers of markers per chromosome. In this case actually BTA27 has the lowest marker density with 147 SNPs. It seems to be likely, that at BTA27 this low SNP density could have been balanced with a low marker average interval (80 kb), because only at BTA30 have significant variations between the different runs been noted. The SNP set of BTA30 is described through an extreme combination of a low SNP density (150) and high average SNP interval (275 kb). Hence it can be suggested, that a high locus density of at least 300 markers with low SNP interval (100 kb) per chromosome will strengthen the order result after different permutations.

The order results of BTA10 and BTA30 indicate that besides the significance of LD, also the nature of LD, SNP density and average SNP interval have to be considered for a successful implementation of the LODE map approach.

The disparities (orphans) between LD data and the bovine assembly used in the present study indicate where the LODE map is expected to give results in conflict to Btau3.1. Hence the CRC integrated map, which provides a more comprehensive bovine map, has been used as an additional basis for evaluation of the LODE map results. Compared to CRC positions, the numbers of orphans is reduced, which results in significantly better order comparisons for BTA1 and BTA15.

5.2.2 Alignment procedure

Besides the possibility to order SNPs based on LD, this study has also shown that it is possible to assign markers to linkage groups and to ordered positions within linkage groups solely based on LD information. Initial principles for this were described by Goddard and Miller (Goddard *et al.* 2005) (Miller *et al.* 2006) and have been conclusively applied to our cattle example.

The initial alignment results of BTA1 produced the expected results as shown by Miller and Hayes (Miller *et al.* 2006), that r^2 values are most useful for dividing closely located SNPs into blocks, i.e. to position current unaligned SNPs on an existing bovine map. In this study this conclusion was also shown with 3D scatter plots, which illustrate the association of already positioned SNPs with unknown SNPs. As expected, D' distances are best for aligning whole scaffolds of SNPs on a single chromosome where only a few previously belonged. In this case, r^2 distances do not demonstrate high correlations between unaligned and known positions. Hence, r^2 distances have been applied more successful to position unknown SNPs on current chromosomes; whereas not all SNPs could have been aligned (e.g. 8 SNPs out of 9 were aligned) concerning BTA1. The SNP which was not assigned to BTA1 was associated with a low maximum r^2 (<0.2) and MAF<0.05.

The alignment results of test batch 2 for the whole genome shows that SNPs with low maximum r^2 (<0.2) to a companion SNP within the chromosome could not have been aligned. In this case it seems likely that especially SNPs with MAF<0.05 are generally associated with low maximum r^2, because out of 22 SNPs only 3 could be aligned. The fact, that also 33 SNPs with MAF>0.05 were not aligned on a single chromosome, indicates that low r^2 values obviously are not only limited to SNPs with MAF <0.05.

The 218 SNPs, which were successfully mapped across the whole genome show that if a SNP with maximum r^2 (>0.4) with another SNP positioned on a single chromosome, the unknown SNP was usually positioned in close proximity (Miller *et al.* 2006).

In this case, it can be expected that, as the maximum r^2 increases, the number of aligned SNP will increase, thereby improving the accuracy of their positioning (Figure 22).

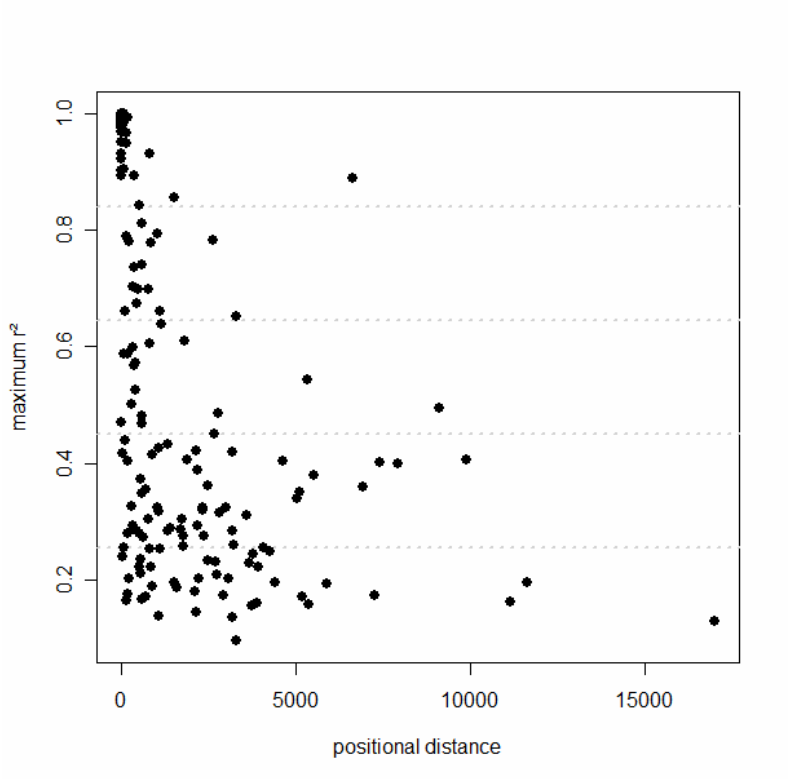


Figure 22 Positional distance (kb) between estimated positions based on maximal r^2 and current Btau3.1 positions.

Concerning the SNP which was placed on two different chromosomes, it has been noted that this SNP has a high maximum r^2 to companion SNPs on both chromosomes (0.998 on BTA11 and 0.993 on BTA13). In this case the SNP can only be positioned based up on the maximum r^2 value, hence on BTA11. However, with a result like this, one must be exceedingly cautious, realizing that the result may reflect aberrant r^2 values or errors in the current map (Miller *et al.* 2006).

The slight variations of the alignment result using LODE map orders instead of Btau3.1 orders for the alignment process shows that obviously the previous order of the markers in the chromosome slightly changes the alignment result. Compared to the alignment run above, no SNP was placed on a different chromosome and an additional two SNPs with $r^2<0.4$ have been placed on two chromosomes.

This result indicates that SNPs placed on different chromosomes are obviously due to low maximal r^2 values or high maximal r^2 on multiple chromosomes and confirms that the wrong placed SNP above is obviously due to aberrant alignment. The application of LODE map positions to calculate the positions of "unknown" SNPs has not resulted in significant differences compared to the use of Btau3.1 positions.

The alignment run of test batch 3 shows that with "low quality" SNPs particularly the numbers of SNPs assigned on multiple chromosomes have been increased. These SNPs assigned on multiple chromosomes were generally associated with unusual high maximum r^2 values. These unusually high r^2 values were normally seen in SNPs which were genotyped only in a few bulls. As expected, using these SNPs with low statistical significance (sample size) of LD has caused a lot of errors in the alignment and ordering procedures, as well as in the computation of the SNP positions, which is shown on an extreme positional distance over 40 Mb at 4 SNP positions (Figure 23).

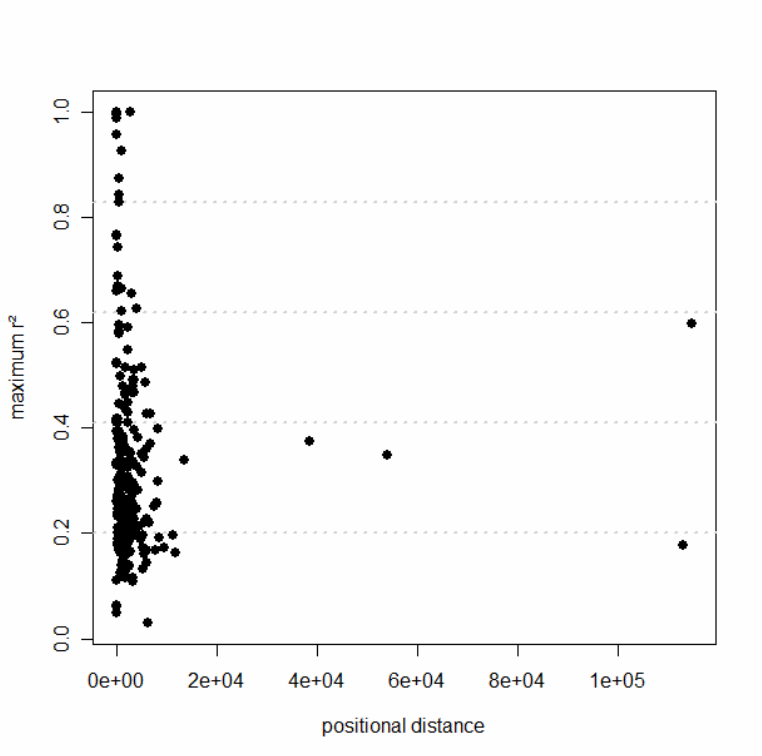


Figure 23 Positional distance (kb) between estimated positions based on maximal r^2 and computed CRC positons.

The high accordance between the positioned SNPs with the LODE and CRC strategies, proofs the quality of both methods has been illustrated. Concerning the 6 SNPs placed on different single chromosomes compared to the CRC positions, there is additional evidence where the alignment based on LD gives a different result to current mapping strategies as already shown at test batch 2.

The alignment result of the 4 SNPs placed on two chromosomes indicates that in some cases the LODE map approach does not give a clear result, because all 4 SNPs have a clear higher r^2 value on one chromosome compared to the other assigned one. In such specific cases it can be suggested that it can be useful to include the maximum r^2 values in the decision of the alignment.

The alignment results of 160 currently unaligned SNPs show that the LODE map approach could be used as a superior tool to position additional SNPs on a current bovine map. Due to the high range of the r^2 values (from 0.12 to 1) of the aligned SNPs, only for about half of the SNPs a high accuracy of the position can be expected, as already shown in test batch 2.

This study shows that ordering and positioning SNPs by exploiting LD may be particularly effective in dairy cattle. While the present study concentrated on the Australian dairy cattle population, it can be predicted that this kind of genetic mapping may also work in other populations where a similar structure of LD can be expected. Such populations include other dairy (Farnir *et al.* 2000) and beef (Odani *et al.* 2006) cattle, commercial pig breeds (Nsengimana *et al.* 2004), horse populations (Tozaki *et al.* 2007) and commercial chicken populations (Heifetz *et al.* 2005). Within the human population the genome-wide use of the LODE map can only be applied to sub-populations in isolated regions (Varilo *et al.* 2003), where there is a similar extend of LD as in the cattle data set.

5.3. The LODE Map strategy

Based on the results of the present study, the recommended strategy to create a LODE map is as follows. Start with a random sample of several thousand SNPs, run SPIN with D' to identify blocks and locus order within blocks. Then prepare the remaining SNPs in sets of, say 1000, add them to each block separately and run SPIN with r^2. This will position most of the unaligned SNPs within a block, and will also indicate which (if any) blocks need to be combined.

In essence, this study shows that it is now possible to create a locus-order map especially in populations showing a high level of LD, due relative low effective population size and/or artificial selection. The five steps involved are:

1. Discover SNPs using the new generation of high-capacity low-cost sequencing technologies;
2. Genotype at least several thousand individuals for all discovered SNPs;
3. Estimate LD for all pair-wise combinations of SNPs;
4. Check the data set concerning the significance and nature of LD;
5. Create a LODE map using the algorithms Side to Side and Neighborhood.

6. CONCLUSION

6.1. GENERAL STATEMENTS on the UTILITY of a Bovine LODE MAP

The results of the current LODE map are unique in providing independent evidence of orders and positions of SNPs that coincide with the current Bovine map. Hence both hypotheses considering the ordering of chromosomes and positioning of unknown SNPs have been confirmed.

Unfortunately with the current developmental stage of the LODE map, a general utility of the procedure for other species can not be confirmed at present. It has already been mentioned that the procedures to create a LODE map are obviously very dependent on the extent and nature of LD of the population sampled. As the nature of LD varies considerably between and within populations, it is impossible to transfer this result to other species. However, it can be predicted that this kind of map will work on most of the chromosomes in populations with a similar extent of LD compared to the Australian Dairy Cattle population.

The method presented in this thesis introduces a new age of genetic mapping, where LD may be used as a completely independent mapping tool, allowing the alignment and positioning of SNPs. Using this new kind of genetic map resulted in better agreement to a consensus map (CRC). Hence it can be predicted, that the additional use of LD will help to improve the quality of current genome maps by:

- Strengthening the positions of already aligned SNPs with LD information;
- Providing independent positions for unaligned SNPs without sequencing;
- Revealing wrongly positioned SNPs (errors in the current map);
- Repositioning SNPs with putative knowledge of the location;
- Using LD as supporting evidence concerning the positions of “Problem SNPs”.

7. REFERENCES

Alexander L.J., Smith T.P., Beattie C.W. & Broom M.F. (1997) Construction and characterization of a large insert procine YAC library. *Mamm. Genome* **8**, 50-1.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-402.

Ardlie K., Kruglyak L. & Seielstad M. (2002) Patterns of linkage disequilibrium in human genome. *Nat. Rev. Genetics* **3**, 299-309.

Barrett J.C., Fry B., Maller J. & Daly M.J. (2005) Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5.

Ben-Dor A., Chor B. & Pelleg D. (2000) RHO---Radiation Hybrid Ordering. *Genome Research* **10**.

Bishop M.D., Kappes S.M., Keele J.W., Stone R.T., Sunden S., Hawkins G.A., Toldo S.S., Fries R., Grosz M.D., Yoo J. & Beattie C.W. (1994) A genetic linkage map for cattle. *Genetics* **136**, 619-39.

Broom M.F. & Hill D.F. (1994) Construction of a large-insert yeast artificial chromosome library from sheep DNA. *Mamm. Genome* **5**, 817–9.

Buetow K.H. & Chakravarti A. (1987) Multipoint gene mapping using seriation General methods. *Am J Hum Genet* **41**, 180-8.

Cavalli-Sforza L.L. (1998) The DNA revolution in population genetics. *Science* **14**, 60 - 5.

Christof T., Jünger M., Kececioglu J., Mutzle P. & Reinelt G. (1997) A branch-and-cut approach to physical mapping of chromosomes by unique probes. *Journal of Comp. Biology* **4**, 433-47.

Collins A., Lonjou C. & Morton N.E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Medical Sciences* **96**, 15173-7.

Cornell U. & IPGRI (2003) Measures of Genetic diversity. In: *Genetic diversity analysis with molecular marker data: Learning module*.

Cox D.R., Burmeister M., Price E.R., Kim S. & Myers R.M. (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Sciene* **250**, 245-50.

Craig G., Nizetic D., Hoheisel J.D., Zehenter G. & Lehrach H. (1990) Ordering of cosmid clones covering the Herpes simplex virus type I (HSV I) genome: a test case for fingerprinting by hybridization. *Nucl. Acids Res.* **18**, 2653-60.

Crawford A.M., Dodds K.G., Ede A.J., Pierson C.A., Montgomery G.W. & Garmonsway H.G. (1995) An Autosomal Genetic Linkage Map of the Sheep Genome. *Gentetics* **140**, 703-24.

Cuticchia A., Arnold J. & Timberlake W. (1992) The use of simulated annealing in chromsome reconstruction experiments based on binary scoring. *Genetics* **132**, 591-601.

Dawson E. (1999) SNP maps: more markers needed? *Molecular Medicine Today* **4**, 419 - 20.

Edwards A. (1963) The measure of association in a 2x2 table. *J. Roy. Stat. Soc. A.* **126**, 109-14.

Falk C.T. (1989) A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. In Elston, Spence, Hodge and MacCluer (ed.), Multipoint mapping and linkage based upon affected pedigree members. *Genetic Workshop* **6**, 17-22.

Faraut T., Givry S.d. & Chabrier P. (2007) A comparative genome approach to makrer ordering. *Bioinformatics* **23**, e50-6.

Farnir F., Coppieters W., Arranz J.J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D. & Georges M. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**, 220-7.

Fung H.C., Scholz S., Matarin M., Simon-Sanchez J., Hernandez D., Britton A., Gibbs J.R., Langefeld C., Stiegert M.L. & Schymick J. (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *The Lancet Neurology* **5**, 911-6.

Garey M.R. & Johnson D.S. (1979) Computer & Intractability: A Guide to the Theory of NP-Completeness. *W H Freeman*.

Goddard M.E., T.H.E & Meuwissen (2005) The use of linkage disequilibrium to map quantitative trait loci. *Aust. J. Exp. Agric.* **45**, 837-45.

Haldane J.B.S. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. *Genetics* **8**, 299-309.

Hästbacka J., Chapelle A.d.l., Kaitiala I., Sistonen P., Weaver A. & Lander E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat. Genet* **2**, 204-11.

Hayes B., Chamberlain J. & Goddard M.E. (2006) Use of linkage markers in linkage disequilibrium with QTL in breeding programs. In: *Proc. 8th World. Congr. Genet. Appl. Livest. Prod.*, Belo Horizonte, Brazil.

Heber S., Stoye J., Hoheisel J. & Vingron M. (2000) Contig Selection in Physical Mapping. *German Cancer Research Center* **1**, 155-64.

Heifetz E.M., Fulton J.E. & O`Sullivan N. (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetis* **171**, 1173-81.

Hill W. & Weir B. (1994) Maximum likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* **54**, 705-14.

Humphreys K. (2007) Measures of LD. *Department of Medical Epidemiology and Biostatistics*

Jeffs B., Negrin C.D., Graham D., Clark J.S., Anderson N.H., Gauguier D. & Dominiczak A.F. (2000) Applicality of a speed congenic strategy to dissect blood pressure quantitative trait loci on rat chromosome 2. *Hypertension* **35**, 179-87.

Joseph Z.B., Gifford D.K. & Jaakkola T.S. (2001) Fast optimal leaf ordering for hierachical clusterin. *Bioinformatics* **17**, S22 - S9.

Khatkar M.S., Collins A., Cavanagh J.A.L., Hawken R.J., Hobbs M., Zenger K.R., Barris W., McClintock A.E., Thomson P.C., Nicholas F.W. & Raadsma H.W. (2006) A First-Generation Metric Linkage Disequilibrium Map of Bovine Chromosome 6. *Genetics* **174**, 79-85.

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., Nicholas F.W. & Radsmaa H.W. (2007) A primary assembly of a bovine haplotype map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle. *Genetics* **176**, 763-72.

Korenberg J.R., Yang-Feng T., Schreck R. & Chen X.N. (1992) U sing fluorescence in situ hybridization (FISH) in genome mapping. *Trends in Biotechnologies* **10**, 27-32.

Laan M.J.v.d. & Pollard K.S. (2003) A new algortihm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statitstical Planning and Inference* **117**, 275 - 303.

Lee W.K., Padmanabhan S. & Dominiczak A.F. (2000) Genetics of hypertension: from experimental modes to clinical applications. *Jouranl of Human Hypertension* **14**, 631-47.

Liu B.-H. (1998) *Statistical genomics : linkage, mapping and QTL analysis*. CRC Press, New York.

Maniatis N., Collins A., Gibson J., Zhang W., Trapper W. & Morton N.E. (2004) Positional cloning by linkage disequilibrium. *Am J Hum Genet* **74**.

Mayraz G. & Shamir R. (1999) Construction of physical maps from olignucleotide fingerprints data. *Proc. RECOMB 99*.

Mester D.I., Ronin Y.I. & Korostishevsky M.A. (2006) Multilocus consensus genetic maps (MCGM): Formulation, algorithms and results. *Computional Biology and Chemistry* **30**, 12-20.

Mester D.I., Ronin Y.I. & Nevo E. (2004) Fast and high precision algorithms for optimization in large-scale genomic problems. *Computional Biology and Chemistry* **28**.

Miller L.D. & Pekny F.J. (1991) Exact Solution of Large Asymmetric Traveling Salesman Problems. *Science* **251**, 754-61.

Miller S.P., Hayes B. & Goddard M.E. (2006) Positioning Single Nucleotide Polymorphisms on an existing bovine map using a genetic algorithm and estimates of linkage disequilibrium. p. 4. University of Guelph, Guelph.

Montanaro V., Casamassimi A., D'Urso M., Yoon J.Y., Freije W., Schlessinger D., Muenke M., Nussbaum R.L., Saccone S., Maugeri S., Santoro A.M., Motta S. & Valle G.D.a. (1991) In situ hybridization to cytogenetic bands of yeast artificial chromosomes covering 50% of human Xq24-Xq28 DNA. *Am J Hum Genet* **48**, 183-94.

Morton N.E. (1955) Sequential fat for the detection of linkage. *Am J Hum Genet* **7**.

Morton N.E., Zhang W., Taillon-Miller P., Ennis S., Kwok P.-Y. & Collins A. (2001) The optimal measure of allelic association. *Medical Sciences* **98**, 5217-21.

Mott R., Grigoriev A., Maier E., Hoheisel J. & Lehrach H. (1993) Algorithms and software tools for odering clone libaries: application to the mapping of the genome of Schizosaccharomyces pombe. *Nucl. Acids Res.* **21**, 1965-74.

Nei M. & Roychodhury A.K. (1988) Human Polymorphyic Genes: World Distribution. *Oxford University Press*.

Nsengimana J., Baret P., Haley C.S. & Visscher P.M. (2004) Linkage disequilibrium in domesticated pig. *Genetics* **166**, 1395-404.

Nsengimana J. & Baret P.V. (2004) Linkage disequilibrium and the genetic distance in livestock populations: the impact of inbreeding. *Genet. Sel. Evol.* **36**, 281-96.

Odani M., Narita A. & Watanabe T. (2006) Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Animal Genetics* **37**, 139-44.

Paschou P., Ziv E., Burchard E.G., Coudhry S., Cintron W.R., Mahoney M.W. & Drineas P. (2007) PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. *PLOS Genetics* **3**, 1672-86.

Pritchard J.K. & Przeworski M. (2001) Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* **69**, 1-14.

Simon-Sanchez J., Scholz S., Fung H.C., Matarin M., Hernandez D., Gibbs J.R., Britton A., Vrieze F.W.d., E. P. & K. G.H. (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozyosity and cell-line induced alterations in normal individuals. *Huma Molecular Genetics* **16**, 1-14

Stallings R.L., Torney D.C., Hildebrand C.E., Longmire J.L., Deaven L.L., Jett J.H., Dogget N.A. & Moysis R.K. (1990) Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl. Acad. Sci.* **USA 87**, 6218-22.

Swinbrune J., Gerstenberg C., Breen M., Aldridge V., Lockhart L., Marti E., Antczak D., Eggelston-Stott M., Bailey E., Mickelson J., Roed K., Lindgren G., Haeringen W.v., Guerin G., Bjarnason J., Allem T. & Binns M. (2000) First Comprehensive Low-Density Horse Linkage Map Based on Two 3-Generation, Full-Sibling, Cross-Bred Horse Reference Families. *Genomics* **66**, 123-34.

Talbot C.J., Cherny S.S., Fulker D.W., Collins A.C. & Flint J. (1999) High-resolution mapping of quantitative trait loci in outbred mice. *Nat. Genet* **20**, 305-8.

Tan Y.D. & Fu Y.X. (2006) A novel method for estimating linkage maps. *Genetics* **173**, 2383-90.

Tozaki T., Hirota K.I. & Hasegawa T. (2007) Whole-genome linkage disequilibrium screening for complex traits in horses. *Molecular Genetics and Genomics* **277**, 663-27.

Tozaki T., Hirota K.I., Hasegawa T., Tomita M. & Kurosawa M. (2005) Prospects for whole genome linkage disequilibrium mapping in thoroughbreds. *Gene* **346**, 127-32.

Trapper W., Maniatis N., Morton N.E. & Collins A. (2003) A Metric Linkage Disequilibrium Map of a Human Chromosome. *Annals of Human Genetics* **67**, 487-94.

Tsafrir D., Tsafrir L., Ein-Dor L., Zuk. O. & Domany E. (2005) Sorting points into neighborhoods (SPIN): data analysis and visiualization by ordeing distance matrices. *Bioinformatics* **21**, 2301-8.

Vaiman D., Schibler L., Bourgeois F., Oustry A., Amigues Y. & Cribiu E. (1996) A genetic linkage map of the male goat genome. *Genetics*, 279-305.

Valdar W., Solberg L.C., Gauguier D., Burnett S., Klenerman P., Cookson W.O., Taylor M.S., Rawlins J.N.P., Mott R. & Flint J. (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet* **38**, 879-87.

Varilo T., Paunio T., Parker A., Perola M., Meyer J., Terwilliger J.D. & Peltonen L. (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Huma Molecular Genetics* **12**, 51-9.

Wang Y., Prade R., Griffith J., Timberlake W. & Arnold J. (1994) ODS_BOOTSTRAP: assessing the statistical reliability of physical maps by bootstrap resampling. *CABIOS* **10**, 625-34.

Weeks D. & Lathrop G. (1995) Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics* **11**, 513-19.

Weeks D.E. & Lange K. (1987) Preliminary ranking procedures for multilocus ordering. *Genomics* **1**, 236-42.

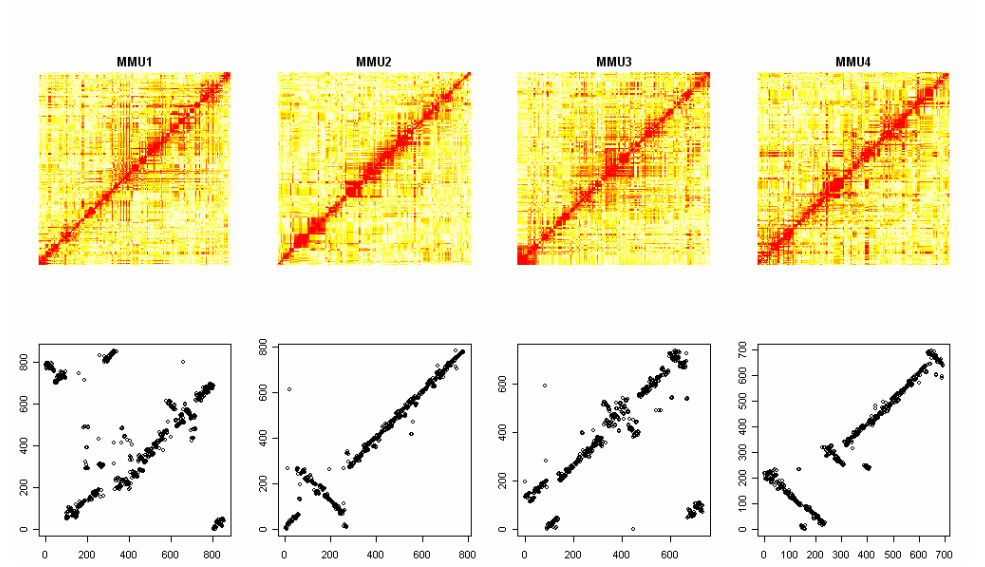
Weiss K.M. & Clark A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* **18**, 19-24.

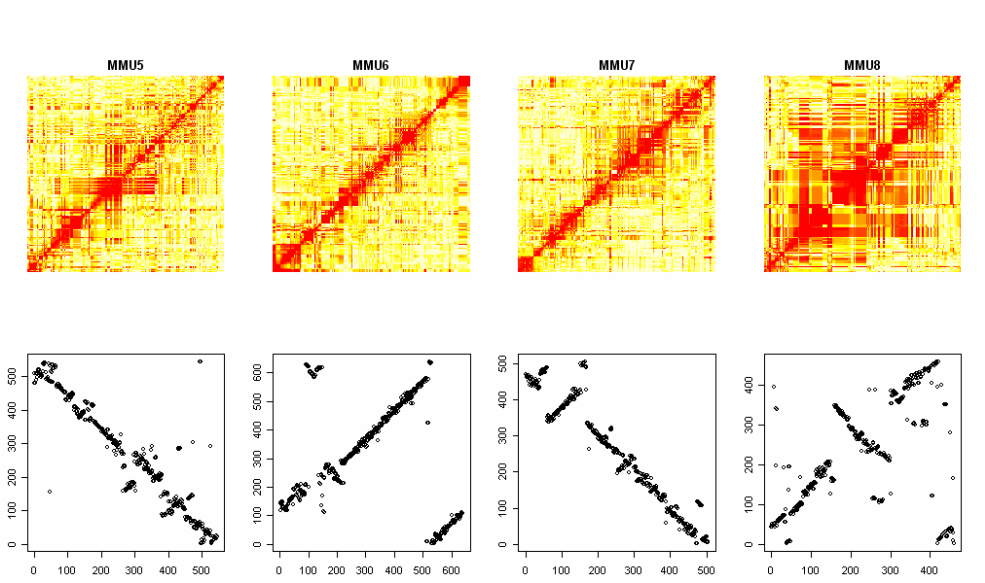
Wilson S.R. (1988) A major simplification in the preliminary ordering of linked loci. *Genet. Epidemiol.* **5**, 75-80.

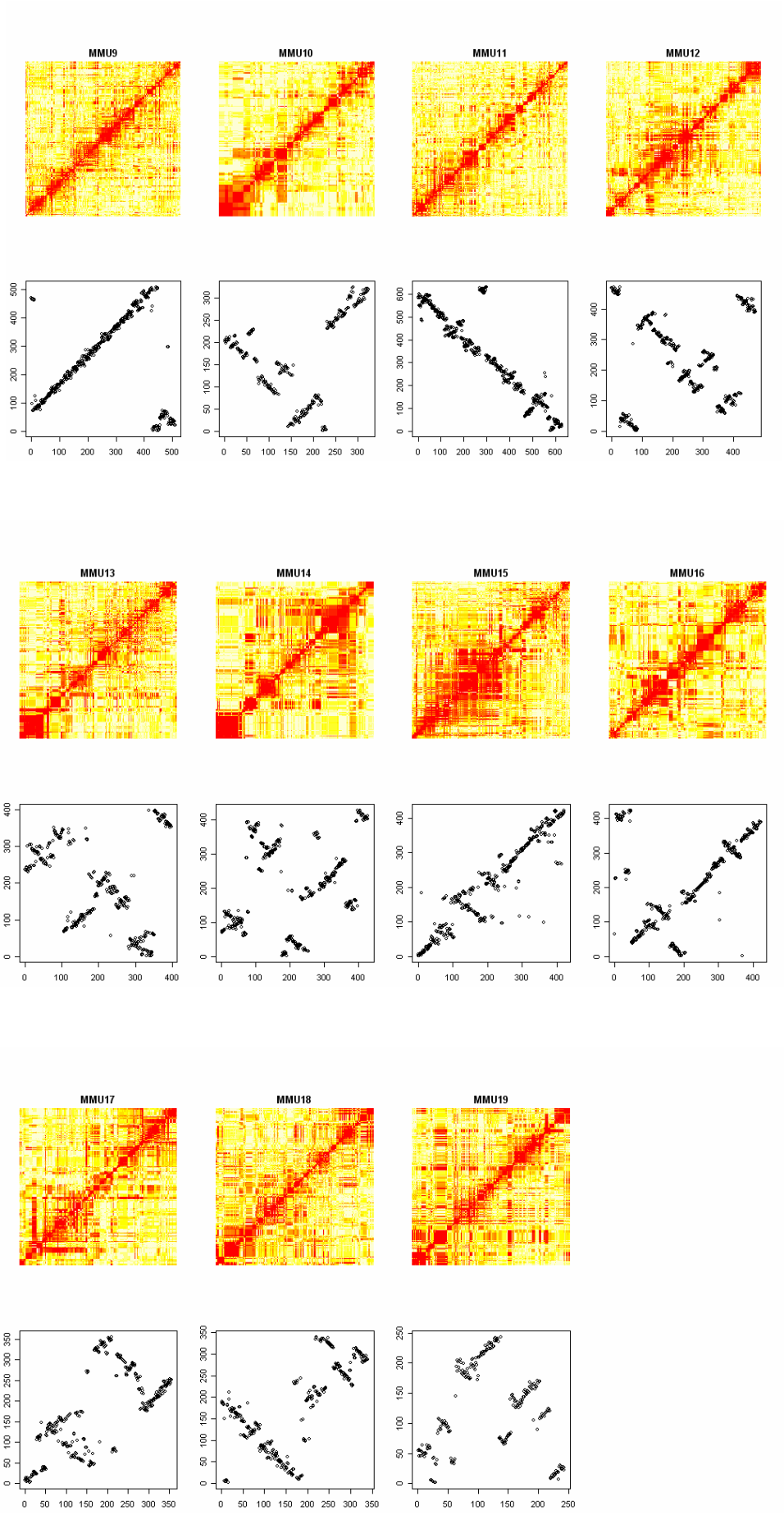
Zhang W., Collins A., Maniatis N., Trapper W. & Morton N.E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl. Acad. Sci.* **99**.

8. APPENDIX

8.1. Whole Murine Genome Results






MMU9
MMU10
MMU11
MMU12
MMU13
MMU14
MMU15
MMU16
MMU17
MMU18
MMU19

8.2. Positions of the 160 aligned SNPs

Assay-ID	LODE Chromosome	LODE Position (kb)	max_r²	Assay-ID	LODE Chromosome	LODE Position (kb)	max_r²
342906	1	1286521	0,21	353402	6	280527	0,27
343199	1	1283105	0,14	353687	6	527494	0,21
343996	1	1286521	0,21	353924	6	527494	0,21
344557	1	1122583	0,44	354086	6	527494	0,22
349106	1	996719	0,34	461862	6	1050611	0,74
349935	1	119735	0,58	463365	6	359025	0,63
351812	1	1225107	0,36	464134	6	857224	0,92
352075	1	241011	0,29	343197	7	25088	0,21
353106	1	264732	0,22	346279	7	556976	0,40
460452	1	1381672	0,23	346960	7	296326	0,25
465161	1	322248	0,28	352334	7	160802	0,57
347371	2	195287	0,43	353542	7	925418	0,56
348500	2	946257	0,25	462866	7	439797	0,98
348926	2	67872	0,14	463279	7	57779	0,32
461764	2	1318334	0,19	463833	7	407513	0,32
342657	3	546029	0,49	464135	7	1004830	0,20
343001	3	1189425	0,12	343166	8	684845	0,22
349886	3	75227	0,13	351687	8	307739	0,39
350884	3	536660	0,21	351688	8	307739	0,39
351243	3	790502	0,44	351829	8	545163	0,29
352215	3	998695	0,23	346242	9	370223	0,23
353429	3	87954	0,37	347662	9	43486	0,23
353844	3	288757	0,31	352151	9	377408	0,48
462065	3	1095969	0,42	352731	9	95959	0,12
464228	3	403001	0,18	352732	9	96118	0,13
464694	3	764504	0,38	353830	9	96118	0,13
466258	3	507755	0,43	462927	9	616287	0,24
350283	4	546359	0,32	464972	9	515031	0,19
351111	4	1122925	0,17	465308	9	900671	0,29
351112	4	1179998	0,30	349822	10	859918	0,19
354389	4	567498	0,39	353556	10	821682	0,29
461443	4	933335	1,00	350130	11	313267	0,43
347190	5	44316	0,13	350781	11	892283	1,00
350289	5	964561	0,20	351473	11	892282	1,00
351011	5	442415	0,18	351474	11	893020	1,00
351228	5	220398	0,41	352106	11	893020	1,00
352946	5	205184	0,12	353704	11	892282	1,00
461482	5	693076	0,52	353705	11	892282	1,00
462628	5	1028407	0,60	461859	11	719197	0,31
462849	5	1142061	0,34	344748	12	17831	0,22
464541	5	338472	1,00	347352	12	17831	0,22
464712	5	597967	0,37	348497	12	166510	0,22
343730	6	280159	0,27	464912	12	23335	0,19
346144	6	280531	0,26	465680	12	788316	0,63

Assay-ID	LODE Chromosome	LODE Position (kb)	max_r²	Assay-ID	LODE Chromosome	LODE Position (kb)	max_r²
343866	13	740812	0,20	463894	20	335542	0,87
345005	13	465544	0,63	464681	20	659931	0,27
350059	13	750608	0,81	346448	21	213447	0,22
353518	13	55666	0,57	347060	21	499611	0,20
461928	13	319882	0,39	350666	21	338915	0,38
463258	13	509894	1,00	462273	21	550210	0,17
343690	14	668328	0,26	465014	21	11869	0,96
343691	14	668328	0,26	346416	22	231960	0,16
347809	14	531731	0,23	349554	22	33455	0,12
463205	14	649583	0,16	350783	22	231936	0,16
463352	14	649583	0,16	353127	22	231936	0,16
463359	14	633210	0,29	461811	22	486461	0,37
463734	14	637649	0,26	464008	22	527576	0,27
347231	15	223408	0,84	464854	22	532835	0,97
349927	15	226752	0,82	343080	23	291776	0,99
351319	15	208158	0,67	464183	23	265919	0,37
351434	15	833245	0,20	464325	23	170347	0,32
353227	15	167187	0,21	465464	23	247363	0,37
461407	15	543672	0,99	465660	23	247994	0,27
463484	15	522586	0,99	345061	24	322986	0,18
463636	15	776646	0,27	354529	24	152772	0,18
350894	16	66722	0,19	464625	24	238310	0,20
353426	16	650873	0,27	352083	25	193908	0,19
464160	16	550359	0,36	353011	25	185641	0,30
352915	17	361342	0,16	343175	26	193950	0,69
353345	17	404137	0,27	463296	27	385439	0,29
463319	17	515721	0,22	347329	28	278562	0,28
346722	18	12955	0,27	353285	28	25953	0,35
349447	18	12955	0,27	343430	29	383391	0,20
352809	18	24397	0,64	344800	29	375118	0,44
462724	18	477681	0,19	344801	29	381162	0,37
343426	19	552595	0,54	346574	29	98513	0,15
349867	19	625140	0,23	342618	30	1038529	0,10
350857	19	552595	0,54	342711	30	1030747	0,23
460522	19	256405	0,32	345227	30	403863	0,38
345063	20	99725	0,23	461068	30	1030747	0,24