Inbreeding determined by the amount of homozygous regions in the genome

Innavl bestemt av mengden homozygoti i genomet

Philosophiae Doctor (PhD) Thesis

Borghild Hillestad

Department of Animal and Aquacultural Sciences

Faculty of Veterinary Medicine and Biosciences

Norwegian University of Life Sciences

Ås 2015

Thesis number 2015:41

ISSN: 1894-6402

ISBN: 978-82-575-1287-3

PhD supervisors

Prof. Gunnar Klemetsdal

Department of Animal and Aquacultural Sciences

Norwegian University of Life Sciences

Box 5003, 1432 Ås, Norway

Prof. John A. Woolliams

The Roslin Institute and Royal (Dick) School of Veterinary Studies

The University of Edinburgh

Box EH25 9RG, Midlothian Easter Bush Campus, Scotland, UK

Prof. Dag Inge Våge

Centre for Integrative Genetics (CIGENE)

Department of Animal and Aquacultural Sciences

Norwegian University of Life Sciences

Box 5003, 1432 Ås, Norway

Prof. Theo H. E. Meuwissen

Department of Animal and Aquacultural Sciences

Norwegian University of Life Sciences

Box 5003, 1432 Ås, Norway

PhD Evaluation Committee

Ass. Prof. Tormod Ådnøy
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences
Box 5003, 1432 Ås
Norway

Prof. Beatrice Villanueva
Animal Breeding Department
National Institute for the Agriculture and Food Research (INIA)
Ctra. de A Coruña Km. 7,5
28040 Madrid
Spain

Dr. Anders Christian Sørensen
Department of Molecular Biology and Genetics –Center for Quantitative Genetics and Genomics
Aarhus University
Blichers Allè 20
8830 Tjele
Denmark

AKNOWLEDGEMENTS

“In order to keep a true perspective of one’s importance, everyone should have a dog that will worship him and a cat that that will ignore him” -Dereke Bruce

This study carried out at the Department of Animal and Aquacultural Sciences at the University of Life Sciences and was funded by the university (project number 11751350). Geno contributed with data, and the project would not be viable without their donations.

I want to thank my supervisors: Professor Gunnar Klemetsdal, Professor John A. Woolliams, Professor Dag Inge Våge and Professor Theo Meuwissen for support and guidance through these three years. Gunnar: I really enjoyed our discussions, both those that contributed to a greater understanding of my project and those that just made us laugh. There have been several hours in our offices, and I have been lucky to get to know a kind and funny man with many good ideas. You have truly deserved every chocolate and licorice we rewarded you. John: When you have the time to spare looking at my work, you always seem to raise it to another level. Thank you for welcoming me to your town, Edinburgh, and for giving me time. Dag Inge can be trusted 100 %. Thank you for answering all my questions, showing an interest even if this topic is a bit on the side of your profession, and for doing a good job within a reasonable time. He is truly a great supervisor, who also turns out to be a good swing dancer. Theo: Thank you for contributing with good suggestions to a solution when I have been stuck, and even reprogramming your software to fit my data.

I also want to thank Geno for being so positive to my work, especially Trygve Roger Solberg and Morten Svendsen. Trygve has been a huge support for me, setting me up with the right people when needed, being on my side when needed and helping me with the professors when their thoughts and visions have wandered way off my PhD’s purpose. Morten: Thank you for sharing your knowledge on the pedigree of Norwegian Red and cracking my problems with a good, old SAS-script when needed.

I had the pleasure to be introduced to runs of homozygosity (ROH) by the team of Johan Sölkner at BOKU in Vienna. Thank you for including me to your group for a couple of days. Thank you Solomon Antwi Boison for teaching me PLINK, and introduce me to genotyping quality controls and ROH definitions.

I am grateful for my brother-in-law Trygve Flathen. He has been helpful programming scenarios to me and introduced me to the world of Linux. He has proven to be quite patient to an impatient PhD-student and sister-in-law. Thank you for being so positive every time I gave you a challenge. Florent Bay has been my R guru, and is always available on e-mail to suggest codes and ways to reach my target, whether it is to create a graph or to measure how big part of the genome that are covered by SNP. I met him at a NOVA course in Latvia, and he has created a folder on his computer with my name on it to store all my R-questions. Thank you also to Harald Grove for giving me an introduction to the world of SNP.

Team Ku (Cow) has been a keystone to me these three years. Cecilie, Bente, Kristine and Katrine: You guys have been the best colleagues ever. Thank you for so many laughs and hours chatting in our office. We have had so many fun trips, lunches and parties. We have been discussing everything from politics and genetics theory to horses, training, men and “Fifty shades of Grey”. You know you are treasured when you after two weeks of absence from the office come back and find the following: (1) Your chair chained by your own bicycle lock to the office desk, (2) dead flies pointedly gathered together in a pile in front of your PC, (3) a hidden keyboard and (4) your wall pictures rearranged. It almost made me cry.

My beloved family! Thanks to my parents Astrid and Thorvald for raising me to the person I am today. My granddad Roar, my two brothers Roar and Erlend and their families and my parents-in-law Ellen and Arne Ivar: Thank you for believing in me. My two wonderful daughters Tonje and Ingunn: thank you for bringing sunshine into my life and for being so patient when Mom had to work long days.

Most of all to my amazing husband Geir: Thank you for being as understanding, supportive, kind, helpful and motivating as you are. This PhD would never been completed without your support. You truly are amazing, and I love you deeply!

Ås, March 2015

Borghild Hillestad

TABLE OF CONTENTS

Paper I-III have individual page numbers

SUMMARY

The main aim of this PhD was to study long homozygote segments present in the genome in Norwegian Red, and find genomic options to measure inbreeding more accurately than from a pedigree database. Prior to the study, runs of homozygosity (ROH) was indicated to be a measure utilizing chromosomal regions identical by descent, thus a good genomic substitute to pedigree. Two dataset were exploited: (1) 384 bulls genotyped with the Illumina HD-panel containing 777K SNP-markers, and (2) 3,289 bulls genotyped with a 54K Illumina BeadChip and/or 25K Affymetrix, with imputations both ways if needed. The pedigree of these two datasets extended as far back as 1875.

Paper I explored how the detection of ROH was affected by SNP density, genotyping quality controls and criteria used to define ROH. It was found that a high SNP density provided increased resolution, fewer false positive ROH, and the possibility to detect shorter ROH. Allowing heterozygote SNP within a ROH as a definition criterion generated false positives. Such a procedure has been common, especially for high SNP densities, to account for genotyping error. Regarding genotyping quality control, pruning for SNP with a low minor allele frequency (MAF) resulted in loss of information. This has been a common procedure working with genotypes in general, but aggravated the quality of the ROH detection.

Paper II compared different approaches to calculate the rate of inbreeding (ΔF) and effective population size (Ne), and studied the effect of SNP density, minimum length of ROH, genotyping quality controls and imputation. Inbreeding coefficients (F) were estimated by utilizing pedigree data (F_{Ped}) and genomic data, both by ROH (F_{ROH}) and observed homozygosity (F_{Hom}). These three inbreeding estimates were regressed on either year of birth or complete generation equivalence (CGE) in a ln(1-F_x) format. The pedigree suffered of a threshold effect, and was not qualified as the best option to measure ΔF and Ne. Observed homozygosity gave the most stable results across SNP density and the best regression fit, accounting for more homozygosity than ROH. By regressing inbreeding coefficients on CGE a better fit was achieved, compared to year of birth. Further, by using a high SNP density and keeping all low MAF SNP, a Ne of 57.5 animals, below a 1/3 of what was obtained by ln(1-F_{Ped}) regressed on year of birth.

Paper III located segments exposed to inbreeding, mapped the rate of inbreeding on a segmental level and searched for selection signatures. By regressing the ln(1-F_{Hom})on CGE, some chromosomes were found to be more inbred than others Chromosomes 5, 6, 14, 20 and 24 had the lowest Ne, ranging between 22.6 and 34.2. Further, positional F_{ROH} was estimated. The highest peaks of inbreeding from ROH were found on chromosomes 1, 5, 7, 14 and 22. Based on logistic regression of ROH status on CGE and ROH-plots, ongoing selective sweeps were located on chromosomes 5, 6, 12 and 24. Footprints like historical sweeps and deserts of missing SNP were also observed.

SAMMENDRAG

Hovedformålet med denne doktorgraden var å studere lange homozygote segmenter i genomet hos NRF, og å finne genomiske metoder som kan måle innavl mer nøyaktig enn ved bruk av slektskapsdatabase. I utgangspunktet var «runs of homozygosity» (ROH) valgt som en egnet og interessant metode for denne studien, fordi den var antatt å oppnå nøyaktige anslag. ROH ble angitt for å være et mål som på lik linje med slektskapsdatabaser utnyttet homosygositet nedarvet fra samme opphav, og dermed en god genomisk erstatning for slektskapsdatabasen. To datasett ble gransket: (1) 384 okser genotypet med Illumina HD-panelet som inneholder 777K SNP-markører, og (2) 3,289 okser genotypet med en 54K Illumina BeadChip og/eller en 25K Affymetrix, med imputering begge veier ved behov. Slektskapsdatabasen til disse to datasettene strakk seg så langt tilbake som til 1875.

Artikkel I gransket hvordan deteksjon av ROH ble påvirket av SNP tetthet, ulike kvalitetskontroller av genotyping og kriterier brukt til å definere ROH. Det ble erfart at en høy SNP-tetthet førte til en mer detaljert deteksjon, en stor andel tidligere feilbestemte ROH forsvant, og det ble mulig å finne ROH av kortere lengder. I tillegg ble det konkludert med at å tillate en heterozygot SNP innenfor et ROH som et definisjonskriterium genererte falske positiver. En slik fremgangsmåte har vært vanlig for å kunne ta hensyn til genotypefeil. Ved preparering av genotypedata, viste det seg at å fjerne SNP med en lav allelfrekvens (MAF) resulterte i tap av informasjon. Også dette har vært et vanlig preparasjonssteg generelt ved analyser av genotyper, men vil i denne sammenhengen forringe kvaliteten på ROH deteksjonen.

Artikkel II sammenlignet ulike tilnærminger for å beregne innavlsrate (ΔF) og effektiv populasjonsstørrelse (Ne), og studerte effekten av SNP tetthet, genotype kvalitetskontroll og imputering. Innavlskoeffisienter ble estimert ved å benytte stamtavle data (F_{Ped}) og genomiske data, både fra ROH (F_{ROH}) og observert homosygositet (F_{Hom}). De tre innavlsestimatene ble regresset i et ln(1-F_x)-format på fødselsår eller antallet komplette generasjoner med stamtavle det var mulig å spore tilbake hos dyret (CGE). En terskeleffekt ble funnet på F_{Ped}, og stamtavle ble derfor ikke regnet som den beste informasjonskilden for å måle ΔF og Ne. Observert homosygositet ga mer stabile resultater på tvers av SNP-tetthet og bedre regresjon, fordi den tok hensyn til mer homosygositet enn ROH. Generelt gav CGE bedre regresjoner enn fødselsår ved en høyere R^2-verdi. Ved å bruke en høy SNP tetthet og beholde alle SNP med lav MAF, ble det beste

estimatet av ΔF oppnådd. Dette resulterte i en Ne av 57,5 dyr, under en 1/3 av det som ble oppnådd ved ln (1-F_{Ped}) regresset på fødselsår.

Artikkel III kartla segmenter på genomet som var utsatt for innavl, ved å definere graden av innavl på et segmentalt nivå og å finne seleksjonssignaturer. Ved regresjon av individuelle F_{Hom}-verdier regresset på CGE, ble flere kromosomer funnet å ha en høyere ΔF enn andre. Hos NRF hadde kromosomene 5, 6, 14, 20 og 24 den laveste Ne, som strakk seg fra 22.6 og 34.2 dyr. Videre ble posisjonelle F_{ROH}-verdier estimert. De segmentene med høyest F_{ROH}-verdier befant seg på kromosomene 1, 5, 7, 14 og 22. Ved hjelp av logistisk regresjon av F_{ROH} på CGE og ROH-plott ble det avdekket «selective sweeps» på kromosomene 5, 6, 12 og 24. Fikserte områder og ørkenområder uten SNP ble også observert.

ABBREVIATIONS

BTA – Bos Taurus Autosome

ΔF – Rate of Inbreeding

F – Individual Inbreeding Coefficient

GEBV – Genomic Estimated Breeding Values

G-matrix – Genomic matrix

GS – Genomic Selection

HWE – Hardy-Weinberg Equilibrium

IBD – Identical by Descent

IBS – Identical by State

LA – Linkage Analysis

LD – Linkage Disequilibrium

MAF – Minor Allele Frequency

Ne – Effective Population Size

ROH – Runs of Homozygosity

SNP – Single Nucleotide Polymorphism

LIST OF PAPERS

The following papers are included in the thesis, and will be referred to by their roman numbers.

Paper I:

Detection of runs of homozygosity in Norwegian Red: Density, criteria and genotyping quality control

Borghild Hillestad, John A. Woolliams, Solomon A. Boison, Harald Grove, Theo Meuwissen, Dag Inge Våge, Gunnar Klemetsdal

Paper II:

Estimating rate of inbreeding and effective population size using genomic data in Norwegian Red

Borghild Hillestad, John A. Woolliams, Theo Meuwissen, Dag Inge Våge, Gunnar Klemetsdal

Paper III:

Screening for selection signatures in Norwegian Red

Borghild Hillestad, John A. Woolliams, Solomon A. Boison, Dag Inge Våge, Gunnar Klemetsdal

GENERAL INTRODUCTION

In genetics, one of the phenomena associated with inbreeding is inbreeding depression, which is synonymous with increased risk of homozygous recessives (Lynch and Walsh, 1998). The corresponding effect are an aggravated score of the phenotype, because the expression of dominance is reduced. The most critical traits subjected to inbreeding depression are those related to fitness where dominance is considered to be more expressed, i.e. traits related to reproduction and offspring survival (Lacy, 1997). For such traits, it is important that natural selection override genetic drift that is known to cause large random changes of allele frequencies. When such changes occur, the rate of inbreeding increases and the effective population size decreases. In practical breeding in Norway, it has been recommended to keep ΔF below 0.5 % per generation for a long time. In addition, FAO (1998) has recommended keeping ΔF below 1 % per generation, stating the importance and priority of controlling inbreeding in commercial livestock populations.

Traditionally, ΔF has been determined by individual inbreeding coefficients (F_{Ped}) or pedigree relationships, generated from pedigree or kinship data (Falconer and Mackay, 1996). To obtain an asymptotic ΔF, the pedigree should be deep enough without errors, likely at least five generations. This is far from practice; there will always be some individuals with either a missing or a wrong pedigree, with errors such as a calf registered to the wrong mother or confusion between semen from two bulls. Such errors lead to an underestimated F, followed by an underestimated ΔF. With an industry relying on underestimated inbreeding measures, populations could unintentionally be at enlarged risk.

One alternative to pedigree is to use dense marker maps to calculate F. By measuring all observed homozygosity of an individual, homozygosity identical by state (IBS) is captured, but inbreeding is defined as homozygosity identical by decent (IBD) and not only IBS. To separate homozygosity IBD from homozygosity only IBS, one option is to focus on homozygosity present in clusters as in ROH. ROH is defined as long homozygote segments present in the genome (Broman and Weber, 1999). Homozygosity caused by recent inbreeding tend to occur as longer segments, because recombination during meiosis from one generation to the next has not yet broken up the segments. Similarly, historical inbreeding will occur as shorter segments, because the chromosome has been broken down through repeated meiosis. An individual inbreeding coefficient from ROH (F_{ROH}) is defined as the ratio between the total length of ROH in an individual and the length of the genome covered by SNP markers (McQuillan et al., 2008). In humans, ROH have been used to differentiate between ethnicities. Humans are not much inbred, but our genome consists of many short ROH, suggesting that humans may have been more inbred in ancient times than now. There are also examples of individuals with long ROH and a high level of relatedness in humans as well (Gibson et al., 2006), and McQuillan et al. (2012) found evidence of inbreeding depression using ROH for human height. Different ethnicities with geographically separation have developed different patterns of ROH, indicating that there are different levels of inbreeding from population to population (Kirin et al., 2010).

The development of SNP chip technology has made it easy to generate large numbers of genotypes per individual. For human genotyping, the densities of the most common chips range between 600K (e.g. Axiom Genome-Wide Human EU and Axiom Genome-Wide ASI) and 2,500K (HumanOmni2.5-8) (Ha et al., 2014). In cattle, the highest density is the Illumina bovine high-density (HD-panel) with a density of 777K, which has dramatically changed the amount of genomic information available compared to lower commonly used chips. A high density is highly desirable, but the cost is correspondingly high. Therefore, cheaper low-density chips, like Affymetrix 25K or Illumina 54K, are commonly used. Lately, new low-density chips have been developed designed as an imputation tool, as the Illumina Bovine low-density (LD) BeadChip with a density of only 7K. Such chips contain markers gaining high imputation efficiency by including markers with: high MAF, even SNP distribution across the genome, high SNP densities at the chromosomal ends, and known haplotypes at the X and Y chromosome as well as the mitochondrial DNA. The variety of densities raises the need to investigate the impact of SNP density and its effect on ROH detection and the potential for imputation to boost the accuracy of detecting ROH when using low-density chips.

Newton-Cheh and Hirschhorn (2005) proposed four characteristics to qualify a marker to be part of a chip: (i) the probability of being functional, (ii) the correlation to expected causal variants (LD), (iii) detected missense variations and (iv) technological considerations. A fifth characteristic may be the functionality of SNP across breeds. If SNP show polymorphism for several breeds, it would increase the commercial advantage to the chip and increase the target audience. Before analysis of genotypes, the genotypes are quality controlled to remove errors. The tradition on quality controls differ from field to field and between different research groups, but the results of the controls will affect the results of the analysis (Edriss et al., 2013; Calus et al., 2014). Call rate, HWE, GenCall score and MAF are elements that are considered in such controls. In GS estimation, pruning of low MAF SNP < 0.05 is common to reduce calculation challenges and increase estimation stability of the remaining SNP, and consequently pruning of low MAF SNP has become a part of the genotyping preparation for ROH (Cole et al., 2009; Kirin et al., 2010; Edriss et al., 2013; Silió et al., 2013). Recently Ferenčaković et al. (2013) chose to rely on call rate and GenCall score only, and not prune for low MAF SNP when detecting ROH. While call rate, HWE and GenCall score can be related to technical errors, the removal of low MAF SNP are population attributes. The chips are species specific and created to fit several breeds. This means that while specific SNP have a high degree of polymorphism in some breeds, they may appear close to or total monomorphic in other breeds. Therefore, there is an interest to find out what effect the pruning of low MAF SNP have to the detection of ROH.

ROH and its qualities are a fairly new discovery, and its definitions remain open. Developed software is limited, and definitions of ROH vary from study to study (Gurgul et al., 2014). The variation is due to several choices: minimum length of a ROH, the allowance of heterozygote or missing SNP within a ROH, average SNP density within a ROH and maximum length of a gap between two SNP within a ROH, to mention some. Some of these constraints also act as genotyping quality controls (e.g. the allowance of heterozygote or missing SNP within a ROH), while others are there to make sure that only two consecutive SNP are not enough to get defined as a ROH (e.g. minimum length). These constraints vary from study to study and make it difficult to compare ROH across projects, and it is of interest to move towards standardizing definitions.

With suitable genomic tools, such as ROH, it is possible to find an improved, genomic substitute to F_{Ped}, to avoid errors and underestimate inbreeding within a population. As both pedigree and ROH intend to focus on the homozygosity IBD, they should in theory both act similar when measuring inbreeding. In a pedigree, there is a base population. These animals may lack known parents, or have been drawn to function as the founders of the population. Because the relationship between the founders either is or have been assumed to be unknown, their inbreeding coefficients are set to zero (Falconer and Mackay, 1996). This way the pedigree stops at a certain point. By increasing the number of generations between the animals of interest to the base population, F_{Ped} will increase. The pedigree of Norwegian Red goes back to the late 1800s and early 1900s, and

F_{Ped} functions as a measure of recent inbreeding. Because short ROH reflects ancient inbreeding, and long ROH recent, it is of curiosity to find how the threshold for minimum length in ROH approaches the pedigree, in case a high threshold for minimum length reflects F_{Ped} better than a low threshold.

By estimating ΔF from individual inbreeding coefficients without the use of pedigree, new possibilities open to wild populations or populations without a pedigree. Inbreeding in wildlife populations have often been measured by Wright's F-statistics using expected heterozygosity (Wright, 1950). This method measures all homozygosity IBS. ROH could accomplish the LD-technique, as LD is less reliable on estimating recent Ne (Corbin et al., 2012). Implementing ROH in inbreeding measures is likely to focus more on homozygosity IBD, removing potential error from the homozygosity that is only IBS. The management and control of populations with a more accurate ΔF or individual F-estimate arrange for a controlled, sustainable and more secure gene conservation program.

When running a breeding program, selection moves segments towards fixation, and favored segments according to the breeding plan will have a greater ΔF than other segments. A population would genetically adapt to environmental changes by selection on new mutations or existing variation, but directional selection could fix either genes or segments, allowing one variant to be the only variant of a gene (Barrett and Schluter, 2008). Opposite to F_{Ped}, F_{ROH} could be a function of position, and each marker would get valued on how it contributes to genomic inbreeding. An elevated F_{ROH} or ΔF on specific segments may indicate selection. By mapping the levels of inbreeding on the genome, it would be possible to detect selection signatures. Thus, it is of interest to develop a positional inbreeding map to maintain a genetic sustainability, control inbreeding and optimize the breeding program.

AIM AND OUTLINE OF THE THESIS

The main objective of this thesis was to utilize dense marker maps to estimate individual inbreeding coefficients and the rate of inbreeding, and to validate whether or not inbreeding is determined more accurately using SNP markers than with pedigree data.

The thesis had three goals:

1. To examine what effect SNP density, genotyping quality control (preferably removal of low MAF SNP) as well as various ROH criteria had on ROH detection.
2. Compare ΔF and Ne estimated from ROH, observed homozygosity and pedigree, and examine the effect of SNP density, minimum lengths to detect ROH, genotyping quality controls and imputation.
3. Map the rate of change of ROH structure on a segmental level and select segments exposed to selection in Norwegian Red.

This thesis was divided into three main parts: Paper I explored how homozygote haplotypes (ROH) appeared and changed according to length and frequency by using different SNP densities, genotyping quality controls and constraints defining a ROH. Paper II estimated inbreeding parameters by the use of molecular and/or pedigree data and explored how these parameters changed when changes were made in either SNP density, minimum length of a ROH, genotyping quality controls or when non-imputed versus imputed data were used. Paper III mapped inbreeding on the chromosome from observed homozygosity, and estimated the rate of change of ROH for each SNP. Visual inspection of ROH distributions over time were also used to discriminate between ongoing and historical selective sweeps.

GENERAL DISCUSSION

This thesis has (i) tested the quality control procedures applied on genotyping data ahead of ROH analysis, (ii) explored the criteria set to define ROH, (iii) established a new theoretical method to measure ΔF and Ne and (iv) mapped positional inbreeding across the genome. The detection of ROH was highly influenced by genotyping quality controls, criteria made for identification of ROH and SNP density. A high SNP density improved the estimates of ROH and provided a higher resolution. By moving from low to high SNP density, several criteria used to define ROH became redundant. However, to avoid false positives it was found of great importance to keep only strictly homozygous segments and not allow heterozygous SNP within a ROH. Pruning of low MAF SNP contributed to loss of information. Estimating Ne and ΔF by using either observed homozygosity or ROH gave more accurate results than from pedigree as the F_{Ped}-values suffered of a threshold effect. Preference was given to observed homozygosity over ROH because it produced stable results of ΔF across SNP densities. ROH gained more from a high density, but produced results intermediate to those from observed homozygosity and pedigree in all densities. ΔF was best estimated when $\ln(1-F_{Hom})$ was regressed on CGE, rather than by year of birth, and resulted in a Ne of 57.5 animals, below 1/3 of what was obtained by $\ln(1-F_{Ped})$ regressed on year of birth. By increasing minimum length of ROH, the quality of the inbreeding measures were set back at a lower density level, and impaired the ROH detection. Imputation without utilizing pedigree information may also have caused additional errors. ROH was found to be an effective screening method when searching for selection signatures without the use of any phenotypes. Norwegian Red had a variable Ne across chromosomes compared to total, average genomic Ne. Selection signatures became visible by logistic regressing positional statuses of ROH on time, showing five segments under ongoing selective sweeps on chromosome 5, 6, 12 and 24.

Animals

Conclusions of a study will always be questioned by the adequacy of the sample. We had acess to two sources of data: (i) 3,289 Norwegian Red bulls genotyped with the Affymetrix 25K and/or the Illumina Beadship 54K, with or without imputation both ways, resulting in a 48K density after quality controls, and (ii) 384 Norwegian Red bulls genotyped with the Illumina HD-panel 777K, leaving 708K after quality controls. The animals with the 48K genotypes were a sample of young Norwegian Red test bulls, born between 1964 and 2009. The animals genotyped with the HD-panel consisted of highly selected breeding bulls (elite bulls), born between 1971 and 2004. Therefore, though 48K-animals were at a closer level to the population mean than the elite bulls, neither of the datasets were random samples of the population.

For elite bulls, a higher proportion of this sample consisted of imported animals compared to the population mean. Norwegian Red has been a synthetic population for a long time with the philosophy of importing the best material. Import of animals contribute to an increase of genetic variation, but might also have contributed to an underestimated F_{Ped}, dependent on the quality of their pedigree data.

In this project, the best accuracy was achieved from the HD-panel despite the lower number of animals. In Paper I it was revealed that a low SNP density gave imprecise results as in false positives and less detected ROH. Paper II showed that even though the animals with the 48K genotypes were a closer fit to the population mean and had 5 times as many animals than the HD-panel group, the estimates from this group based on pedigree were similar to the same estimates from the HD-panel group. This indicates that the animals genotyped with the HD-panel worked well as candidates for the population, even though they were not randomly chosen.

ROH as an inbreeding measure across species

Besides cattle, inbreeding studies using ROH have been performed both in humans (Pemberton et al., 2012) and in pigs (Silió et al., 2013). Cattle, the species of this thesis, was domesticated for approximately 10,500 years ago in the Near East (Bollongino et al., 2012). Since then, selection has been carried out in cattle, either systematic or unsystematic. Norwegian Red has been under a systematic selection program since the early 1900s. Because of domestication and systematic breeding, ROH appear in different lengths. Paper III showed how the dataset of 381 bulls contained ROH with lengths ranging between 0.5 up to 58.7 Mb, and the longest ROH was approximately equal to half a chromosome. Even though outbreeding is more common in humans than in cattle, resulting in ROH with a lower average length, ROH seem to be a tool detecting inbreeding also in humans (McQuillan et al., 2008; Pemberton et al., 2012). Mammalian genomes in general vary broadly in physics and appearance, but the majority of mammalian genes are orthologous, meaning that they arose before the species were developed and are therefore present in several species (Gibbs et al., 2004; Elsik et al., 2009). Therefore, it should be possible to use ROH in all mammals, despite their differences. To locate ROH in a species, the following criteria must hold: (i) The genome used must have been sequenced; if using SNP chips (ii) the physical location of the SNP must be known; and (iii) low MAF SNP should not be removed. Also, to achieve good and reliable results a chip of high density is recommended, and a great effort and considerations should be put into the genotyping quality controls and the criteria set to identify ROH. When detecting ROH in species other than cattle, the recommendations of criteria found in this project could be used as a starting point to define ROH, but should be adjusted to the specific species if needed.

The value of pedigree information

The thesis showed that ΔF estimates from pedigree might suffer from insufficiencies in data; on the bull side, as mentioned, but also through dams as herd recording was only complete for cows born 1978 onwards. In this situation, it is logical that genomic data supplies more information. Paper II gave a good picture on how inclusion of both pedigree and genomic data provided more accurate estimates compared to separate analyses: Inbreeding was estimated from ROH, observed homozygosity and pedigree, and the results were compared. We demonstrated that ΔF and Ne were best estimated from $\ln(1\text{-}F_{Hom})$ regressed on CGE, where $\ln(1\text{-}F_{Hom})$ is based on individual genotypes and CGE is calculated from the pedigree of the animal. In populations with non-overlapping generations and a complete pedigree back to the base, regressing on CGE would not have any value, and regressing on year of birth would be needed. This is the option for wild populations, that need to be further studied and compared.

A combination of genomics and pedigree also seemed to be an advantage in imputation. For an imputation tool to build haplotypes, the tools available are either relying on both genotypes and pedigree as in LDMIP (Meuwissen and Goddard, 2010) or AlphaImpute (Hickey et al., 2012), or rely on genotypes through LD, as in Beagle (Browning and Browning, 2007). Paper II pointed out the possibility of imputation without using a pedigree contributing to error when estimating rate of inbreeding from imputed datasets. Daetwyler et al. (2011) also found an advantage of comparing relatives when imputing genotypes: computer time and error rates were reduced, because animals were compared to relatives and not the whole dataset. This once again suggests that pedigree pushes the genotypes to better estimates.

Potential use of genomic inbreeding measures

For traits with non-additive genetic effects, genomic inbreeding would be suited to estimate inbreeding depression or heterosis. Martinsen et al. (2013) used F_{Ped} to show negative effects of inbreeding on milk and fertility traits in Norwegian Red, and Christensen et al. (1996) reported in an early study a negative effect of inbreeding on growth in pigs by studying 21 marker loci. By substituting F_{Ped} with F_{Hom} or F_{ROH} inbreeding depression or heterosis would likely be detected as long as effects of dominance and epistasis are present for the trait. Further, Luan et al. (2014) showed that a G-matrix built from ROH could give more accurate GEBVs than when building G-matrices from LA or IBD information, showing how ROH may give SNP wise additive estimates of breeding values. Also, in paper III chromosomal F_{Hom}-values and positional F_{ROH}-values on each SNP were calculated. By estimating inbreeding depression based on either chromosomal F_{Hom}-values or positional F_{ROH}-values inbreeding depression could be detected on a chromosomal or a segmental level. By knowing where on the genome each animal are inbred, the mating options would radically change.

CONCLUSIONS

The main findings of this thesis were:

The detection of ROH was highly influenced by genotyping quality controls, criteria made for identification of ROH and SNP density:

- A high SNP density improved the estimates of ROH and improved the resolution.
- By moving from low to high SNP density, several criteria used to define ROH became redundant, except the allowance of heterozygote SNP within a ROH. By allowing heterozygote SNP in a ROH when the density was increased, false positive ROH was created instead of adjusting for genotyping errors.
- Pruning of low MAF SNP contributed to loss of information.

When comparing F-values from pedigree, observed homozygosity and ROH, the rate of inbreeding and effective population size were best estimated by regressing ln(1-F_{Hom}) on CGE using a 708K density:

- F_{Ped}-values suffered of a threshold effect and did not manage to distribute the actual genetic variation very well. Thus, too much weight was allocated to animals with high inbreeding coefficients in the regression.
- Preference was given to observed homozygosity over ROH because it produced stable results of ΔF across SNP densities and had a better regression fit with a higher R^2 than ROH.
- ROH performed better with a high rather than a low SNP density, and produced results intermediate to those from observed homozygosity and pedigree.
- In this population CGE was found to be a better explanatory variable than year of birth, as a better regression fit was achieved.
- Imputation programs that do not include pedigree information may fail in detecting homozygosity and should be investigated further.
- The best estimate of Ne for Norwegian Red was 57.5 animals, below 1/3 of what was obtained by ln(1-F_{Ped}) regressed on year of birth.

Regressing ROH statuses on time revealed to be an effective screening method searching for selection signatures without any phenotypes available:

- Norwegian Red had a decreased Ne on several chromosomes compared to total genomic Ne. BTA 5, 14 and 25 were found to be Bonferroni significant with Ne ranging between 22.6 and 34.2.
- The highest values of $F_{j(0.5)}$ were found on chromosome 1, 5, 7, 14, and 22, indicating much homozygosity on these chromosomes
- Selection signatures became visible by logistic regressing of ROH status on time, showing 4 segments being under ongoing selective sweeps in chromosome 5, 6, 12 and 24.

RECOMMENDATIONS

- When working with ROH: Do not prune away low MAF SNP, use a high SNP-density and be careful with how ROH is defined
- Rate of inbreeding and effective population size is best estimated by regressing ln(1-F_{Hom}) on CGE, and alarms us that pedigree based estimates in Norwegian Red may have been overestimated Ne by approximately 300 %. This should be followed up by additional research with more data.
- ROH and possibly observed homozygosity can be utilized to screen for selection signatures.

REFERENCES

Barrett, R. D. H., and D. Schluter. 2008. Adaptation from standing genetic variation. Trends Ecol. Evol. 23: 38-44.

Bollongino, R. et al. 2012. Modern Taurine Cattle descended from small number of Near-Eastern founders. Molecular Biology and Evolution.

Broman, K. W., and J. L. Weber. 1999. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. Am. J. Hum. Genet. 65: 1493-1500.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81: 1084-1097.

Calus, M. P., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal : an international journal of animal bioscience 8: 1743-1753.

Christensen, K., M. Fredholm, A. K. Winterø, J. N. Jørgensen, and S. Andersen. 1996. Joint effect of 21 marker loci and effect of realized inbreeding on growth in pigs. Animal Science 62: 541-546.

Cole, J. B. et al. 2009. Distribution and location of genetic effects for dairy traits. Journal of Dairy Science 92: 2931-2946.

Corbin, L. J., A. Y. H. Liu, S. C. Bishop, and J. A. Woolliams. 2012. Estimation of historical effective population size using linkage disequilibria with marker data. Journal of Animal Breeding and Genetics 129: 257-270.

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Woolliams, and M. E. Goddard. 2011. Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. Genetics 189: 317-U1028.

Edriss, V., B. Guldbrandtsen, M. S. Lund, and G. Su. 2013. Effect of marker-data editing on the accuracy of genomic prediction. Journal of Animal Breeding and Genetics 130: 128-135.

Elsik, C. G. et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science (New York, N.Y.) 324: 522-528.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4 ed. Pearson Education Limited, England.

FAO. 1998. Secondary guidelines for development of national farm animal genetic resources management plans -Management of small populations at risk. In: F. a. A. O. o. t. U. Nations (ed.).

Ferenčaković, M., J. Sölkner, and I. Curik. 2013. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics, selection, evolution : GSE 45: 42-42.

Gibbs, R. A. et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493-521.

Gibson, J., N. Morton, and A. Collins. 2006. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15: 789 - 795.

Gurgul, A. et al. 2014. The application of genome-wide SNP genotyping methods in studies on livestock genomes. Journal of applied genetics 55: 197-208.

Ha, N. T., S. Freytag, and H. Bickeboeller. 2014. Coverage and efficiency in current SNP chips. European Journal of Human Genetics 22: 1124-1130.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. van der Werf, and M. A. Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics, selection, evolution : GSE 44: 9.

Kirin, M. et al. 2010. Genomic runs of homozygosity record population history and consanguinity. PLoS One 5: e13996.

Lacy, R. C. 1997. Importance of genetic variation to the viability of mammalian populations. J. Mammal. 78: 320-335.

Luan, T., X. J. Yu, M. Dolezal, A. Bagnato, and T. H. E. Meuwissen. 2014. Genomic prediction based on runs of homozygosity. Genet. Sel. Evol. 46: 9.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, Mass.

Martinsen, K. H., E. Sehested, and B. Heringstad. 2013. Effects of inbreeding on milk production, fertility, and somatic cell count in Norwegian Red. In: EAAP Annual Meeting, Nantes, France. p 610.

McQuillan, R. et al. 2012. Evidence of Inbreeding Depression on Human Height. Plos Genetics 8: 14.

McQuillan, R. et al. 2008. Runs of homozygosity in European populations. Am J Hum Genet 83: 359 - 372.

Meuwissen, T., and M. Goddard. 2010. The Use of Family Relationships and Linkage Disequilibrium to Impute Phase and Missing Genotypes in Up to Whole-Genome Sequence Density Genotypic Data. Genetics 185: 1441-1449.

Newton-Cheh, C., and J. N. Hirschhorn. 2005. Genetic association studies of complex traits: design and analysis issues. Mutation research 573: 54-69.

Pemberton, T. J. et al. 2012. Genomic Patterns of Homozygosity in Worldwide Human Populations. Am. J. Hum. Genet. 91: 275-292.

Silió, L. et al. 2013. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. Journal of Animal Breeding and Genetics: n/a-n/a.

Wright, S. 1950. Genetical structure of populations. Nature 166: 247-249.

Paper I

Detection of runs of homozygosity in Norwegian Red: Density, criteria and genotyping quality control

Borghild Hillestad, John Arthur Woolliams, Solomon Antwi Boison, Harald Grove, Theo Meuwissen, Dag Inge Våge, Gunnar Klemetsdal

Submitted to Genetic Selection Evolution

Detection of runs of homozygosity in Norwegian Red: Density, criteria and genotyping quality control

Borghild Hillestad[1], John Arthur Woolliams[1,2], Solomon Antwi Boison[3], Harald Grove[1,4], Theo Meuwissen[1], Dag Inge Våge[1,4], Gunnar Klemetsdal[1]

[1]Department of Animal and Aquacultural Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, N-1432 Ås, Norway

[2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

[3]University of Natural Resources and Life Sciences Vienna, Department of Sustainable Agricultural Systems, Division of Livestock Sciences, Gregor Mendel Str. 33, A-1180 Vienna, Austria

[4]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, N-1432 Ås, Norway

Borghild Hillestad borghildhillestad@gmail.com

John Arthur Woolliams john.woolliams@roslin.ed.ac.uk

Solomon Antwi Boison soloboan@yahoo.com

Harald Grove harald.grove@nmbu.no

Theo Meuwissen theo.meuwissen@nmbu.no

Dag Inge Våge daginge.vage@nmbu.no

Gunnar Klemetsdal gunnar.klemetsdal@nmbu.no

Corresponding author: Borghild Hillestad

Abstract

Background. Runs of homozygosity (**ROH**) are long, homozygote segments of an individual's genome, traceable to the parents and might be identical by descent (**IBD**). Due to the lack of standards for quality control of genotyping and criteria to define ROH, Norwegian Red was used to find the effects of SNP density, genotyping quality control and ROH-criteria on the detection of ROH.

Materials and Methods. A total of 384 bulls were genotyped with the Illumina HD-chip containing 777,962 SNP-markers. A total of 22 data subsets were derived to examine effects of SNP density, quality control of genotyping and ROH-criteria. ROH was detected by PLINK.

Results and Conclusions. High SNP density leaded to increased resolution, fewer false positive ROH, and made it possible to detect shorter ROH. Considering the ROH criteria, we demonstrated that allowing for heterozygote SNP could generates false positives. Further, genotyping quality control should be tuned towards keeping as many SNP as possible, also low MAF SNP, as otherwise many ROH will be lost.

Keywords: Runs of homozygosity, SNP density, ROH standards, Low MAF SNP

Background

Runs of homozygosity (**ROH**) are stretches of homozygous segments present in the genome caused by parents transmitting identical haplotypes to their offspring. If two copies of the same ancestral haplotype are passed on to an offspring, homozygosity occurs [1]. Over its length, the frequency of homozygosity depends on the history and the management of the population. The use of the molecular markers in the human data, allowed Broman and Weber to demonstrate the relationship between the length of the homozygous segment and the length of time from the common ancestor. A homozygous segment originating from a more recent ancestor is expected to be longer as there have been fewer opportunities for recombinations to reduce its length. This makes it possible to characterize subpopulations based on the length of the homozygous segments. For instance; human subpopulations that allow cousin marriage tend to have longer average ROH compared to subpopulations that do not allow cousin marriage, because closely inter-related subpopulations contain longer segments compared to outbred subpopulations [2]. Although the proportion of the genome that is homozygous, irrespective of length, can be used as a measure of observed inbreeding, a distinctive feature of ROH has the possibility to distinguish between recent and ancient inbreeding [3]. By looking at the ratio between the total length of ROH in an individual and the length of the genome, an observed inbreeding coefficient (**F_{ROH}**) is created [4].

However this simple idea has debatable issues, primarily around the idea of a haplotype. F_{ROH} is not defined absolutely in the absence of sequence, and typically relies on SNP marker data. Therefore a ROH depends *a priori* on parameters used to define the length of the ROH when it is inferred from markers. These parameters are often associated with the quality control applied to the marker genotypes, and this differs from study to study. A common procedure has been the removal of SNP with minor allele frequency (**MAF**) below a certain threshold; as this has been common in genome-wide association studies (**GWAS**), it has also become accepted as a genotyping quality control in ROH-analysis [5-8]. A justification of this procedure in GWAS has been to avoid SNP whose effect may be sensitive to rogue phenotypes or sub-structures, but an additional purpose is to remove SNP that have been incorrectly genotyped. Whilst the latter is relevant to ROH, the former is not, and hence it remains a question whether removal of low MAF SNP is really necessary for ROH estimation, and if such control measures improve the detection and value of F_{ROH}.

This question becomes more relevant if the primary processing of genotype data is for use in genomic selection or genetic relationship matrix (**G**), for instance by genomic selection (**GS**) [9]. In the context of GS it is common to delete SNP with MAF as high as 0.05 [10]. Other studies like Keller et al. [11] have pruned MAF > 0.05, when using different F coefficients based on SNP to investigate the power for detecting inbreeding depression. Studies such as these highlight the importance of quality controls on the SNP-data designed for different purposes.

The criteria set to define ROH will affect what and how much we detect of clustered homozygosity. It is of interest to find the optimum criteria and to know what gives the most accurate and informative detections in ROH to define inbreeding. Herein, the aims were to examine the effects of SNP density, genotyping quality control (preferably removal of low MAF SNP) as well as various ROH criteria on ROH detection.

Materials and Methods

Detection of ROH in data subsets with different SNP densities for predefined ROH criteria

The impact of SNP-density on the detection of ROH were examined in 384 Norwegian Red bulls genotyped with the Illumina HD-panel. The panel contains 777,962 SNP-markers, covering 2.51 Gb of the 3 Gb large genome, although not all these SNP-markers will be polymorphic in the Norwegian Red. After genotyping, the marker data passed through several stages of quality controls, or genotype editing, to exclude markers on sex-linked chromosomes, call rate per SNP > 90 % (individual SNP score missing if GenCall score < 0.7) and deviation from Hardy-Weinberg ($P > 10^{-6}$) (Table 1). Three animals were deleted for having genotypes for fewer than 95 % of loci. This resulted in the retention of 707,609 SNP, which will be denoted the 708K set.

The 708K set was sequentially pruned to give further nine subsets of data. The first pruning removed every fourth SNP, by physical order, from the 708K set to obtain a subset of 530,706 SNP (denoted 531K set). This procedure was repeated by removing every fourth SNP from the 531K set, to obtain a 398K set, and a further seven times to give the smallest subset (a 53K set). All densities achieved are shown in Table 2.

For each of these sets ROH were identified with PLINK 1.07 [12]. PLINK takes a window of 5,000 Kb and slides it across the genome, determining homozygosity at each window. The identifications of ROH requires specifications of criteria concerned with (i) the minimum number of adjacent homozygous SNP loci to define a run; (ii) the number of heterozygous SNP allowed within a window, which is permitted as they are presumed to be genotyping errors; (iii) the number of missing SNP allowed within a window; (iv) the maximum physical distance between adjacent SNP within a run (maximum gap length); and (v) the minimum density of SNP within a run (average Kb per SNP). These ROH criteria differed according to the SNP-density of the subset used, and are shown in Table 3.

Detection of ROH when altering ROH criteria

First, the effect of allowing one heterozygote SNP per window were examined by generating another subset (708K$_{\mathrm{Alt1}}$) that did not allow for any heterozygote SNP per window (Table 3). Secondly, the effect of applying ROH criteria used for lower SNP density sets was examined by generating three datasets; 708K$_{\mathrm{Alt2}}$, 708K$_{\mathrm{Alt3}}$ and 708K$_{\mathrm{Alt4}}$, that used the same criteria as used for densities of 53-94K, 126K and 168-299K, respectively. Further, the effect of reducing number of missing SNP per window from 3 to 1, otherwise for the same criteria as in 708K$_{\mathrm{Alt1}}$ led forward to set 708K$_{\mathrm{Alt5.}}$ Finally, the effect of increasing the maximum gap length, for the same average SNP density, was examined by use of set 708K$_{\mathrm{Alt6,}}$ while the effect of an increase of the allowed maximum average Kb per SNP relied on set 708K$_{\mathrm{Alt7}}$.

Detection of ROH with varying MAF thresholds

To find what effect removal of low MAF SNP has on ROH detection, two additional subsets were defined based on the 708K set. These were obtained by pruning SNP with MAF < 0.01, resulting in a loss of approximately 14 % SNP and a total of 610,885 SNP (611K$_{\mathrm{MAF}}$). A further subset was obtained by removing SNP with MAF < 0.02; resulting in an additional 2 % of SNP and a total number of 597,454 SNP (597K$_{\mathrm{MAF}}$) (Table 2). In both these datasets, identification of ROH was done as earlier described with criteria given in Table 3. Differences between ROH identified with 708K, 611K$_{\mathrm{MAF}}$ and 597K$_{\mathrm{MAF}}$ were investigated and classified according to chromosomes.

Heterozygosity on a chromosomal level

For the 708K set, average rate of heterozygosity (**Het**) was estimated on each chromosome based by the following equation:

$$\mathrm{Het} = \mathrm{O}\left(\mathrm{Hom}\right) / \mathrm{N}\left(\mathrm{NM}\right) \qquad (1)$$

,where O(Hom) is observed homozygosity and N(NM) is defined as the number of non-missing genotypes.

Results

Variation in SNP-densities and ROH criteria

Minimum number of homozygeous SNP/Kb. With a minimum threshold set both in Kb and in number of SNP, this is directly reflected in the missing pattern of Table 4, e.g. ROH shorter than 2 Mb could not be detected when the criterion set the threshold for minimum length to 2,000 Kb, as for 53K – 94K (Table 3).

SNP density. Across the 10 sets with differing SNP densities, the average number of ROH in an individual differed from 23.2 (53K) to 209 (398K) (Table 4). The maximum number of observed ROH was therefore not found in the densest SNP set, but in the 398K set. The effect of SNP density could be seen within groups: 53K, 71K, 94K and 708K$_{\mathrm{Alt2}}$ sets; 126K and 708K$_{\mathrm{Alt3}}$ sets;

224K, 299K and 708K$_{Alt4}$ sets and the 398K, 531K and 708K sets, where in each of these groups the additional criteria remained constant (Table 3). In principle, with constant additional criteria, using more SNP to detect ROH would be expected to reduce the observed numbers of long ROH and total length of ROH as the additional SNP will help to remove the false positives that may have been identified with the lower SNP density. For the first group and with increasing density, there was observed a redistribution of ROH, from longer to shorter ROH that also reduced the total length (Table 4).

Despite that lower densities were incapable of detecting shorter lengths (< 2 Mb) when other criteria were applied, the effect of increasing density in the 53K, 71K, 94K and 708K$_{Alt2}$ sets was an increased number of ROH detected (Table 4). Since the 53K set contained on average only 88.5 SNP in a 5 Mb window and as much as 15 SNP were required to establish a ROH of length 2 Mb, fewer ROH of lengths between 2Mb and 4Mb were detected with the 53K set than the 94K set. The 94K set had an average of 157.4 SNP in a 5 Mb window, and detected 13.1 ROH between 2 and 4 Mb (cf. 9.8 in the 53K set). Similarly, the 708K$_{Alt2}$, with a coverage of 1,179.3 SNP per window detected 14.4 ROH in the 2-4 Mb category.

The mentioned redistribution of ROH was also seen for the three other groups, but now ROH < 2 Mb decreased in number as the chip became denser and false positives were removed; therefore the high density sets provide better estimation possibilities of shorter ROH than low density sets. Actually, of the 184.1 ROH detected in 708K data, 71 % were found in the shortest category (0.5 – 1 Mb) considered here.

Heterozygous SNP. Another contrast in the SNP density sets (126K cf. 168K of Table 3) was the allowance heterozygote SNP within a ROH. When SNP density increased it was expected that the number of detected ROH of the different ROH groups increased more for short ROH than for long ROH. In the 1-2 Mb category, the number of ROH detected increased by 63.8 % and in the next category (2-4 Mb) the detected ROH increased by 6.9 % (Table 4). However the other densities suggest that the gain in the number of ROH was primarily in false positives. For the 1-2 Mb category the 708K set detected ROH intermediate between the 126K set and the 168K set, but closer to the 126K set. Almost all the additional ROH in the 2-4 Mb category were removed subsequently as being false positives.

Comparison of results for 708K with those for $708K_{Alt1}$ (Table 4) indicates that allowing heterozygotes (in 708K) also added false positives to defined short ROH: by allowing one heterozygote SNP per window, the amount of short ROH (0.5-1 Mb) increased with 46.8 %, while long ROH (8-16 Mb) increased with only 8.3 % (Table 4).This suggests that avoidance of heterozygote SNP are needed to further reduce detection of false positives.

Also in the $708K_{Alt1}$ set, the frequency of short ROH were higher compared to longer ROH (Table 4); the occurrence of ROH in the 0.5-1 Mb category was close to four folds the 1-2 Mb category, clearly illustrated by the cumulative distribution of number of detected ROH by ROH-lengths (Figure 1).

Missing SNP. For an individual, some SNP will be missing. Here, the effect of allowing three missing SNP per window vs only one missing SNP was examined (Table 4: $708K_{Alt1}$ vs

$708K_{Alt5}$), otherwise for the same criteria. The effect was only minor; the number of long ROH had a small tendency to increase with increased number of missing SNP allowed, but did not affect the results much.

Maximum average density and maximum gap length. Maximum average densities of 150 and 50 Kb were compared, and had roughly no effect on the results (Table 4: $708K_{Alt7}$ vs $708K_{Alt1}$). Further, using maximum gap lengths of 1,000 and 250 Kb gave only a minor effect (Table 4: $708K_{Alt6}$ vs $708K_{Alt1}$).

MAF. The two MAF sets $597K_{MAF}$ and $611K_{MAF}$ had ROH criteria identical to the 398K, 531K and 708K SNP sets (Table 3). Both these MAF sets detected fewer ROH than both the 531K and the 708K set, where the major differences appeared at the 0.5-1 Mb category (Table 4). By mapping the loss of short ROH from 708K to $597K_{MAF}$ by chromosome (Table 5), it appeared that the low MAF SNP removed were unevenly distributed: BTA 8, 13 and 14, respectively, lost 30.8, 27.0 and 28.3 % of the total amount of SNP in the chromosome when SNP with MAF < 0.02 were removed compared to the average loss of 15.7 % over the whole genome. When limiting results to short ROH (0.5-1 Mb), the number was unevenly affected by removal of low MAF SNP: BTA 13 and 14 lost 18.6 and 19.7 % of short ROH by pruning for MAF < 0.02, compared to the total average of 8.3 %, suggesting that low MAF SNP are associated with the ROH and/or criteria used. This could be a sign of selection signatures. Further support for selection signatures came from the lowered average rate of heterozygosity on BTA 13 and 14 of 0.343 and 0.341, respectively, relative to a total average of 0.355 (Table 5).

Discussion

There is a need to set standards of the constraints when ROH is used to estimate inbreeding. Because both genotyping quality control and constraints to detect ROH are different from study to study, it is difficult, if not impossible to compare results [13]. In this study we altered on common variables and constraints within SNP density, genotyping quality controls and criteria to detect ROH, where several factors rather gained than removed error.

A higher SNP density improved the resolution, reduced errors by rescaling long ROH to shorter ROH, refusing falsely detected ROH from low densities and by allowing shorter ROH to be detected. When ROH is wanted, it is of great importance to keep as many SNP as possible in order to achieve a picture of how homozygosity is distributed. And by using a high SNP density, more details contributes to a more accurate estimate. There is no doubt that a high SNP density contribute to a more precise estimate of ROH than a low density.

By using a high threshold for minimum length when detecting ROH, massive information on homozygosity were rejected. Short ROH, that are likely to have been exposed to recombination over a long time, relates to a more ancient base than that of the long ROH. Minimum length of ROH of 0.5 Mb was defined in accordance with Ferenčaković et al. [8], to avoid ROH that were more likely arise due to population linkage disequilibrium rather than due to inheritance. There has been speculations whether it would be appropriate to raise the minimum length of ROH in order to capture recent inbreeding and avoid ancient inbreeding that no longer concerns the population, which is why the minimum length has been raised in some studies [14, 15]. When inbreeding were measured by ROH, massive homozygosity were rejected and assumed not to be IBD. Because we do not know if this assumption is correct, and because some of the approved ROH also may not be IBD, we should be careful about removing even more homozygosity by raising the threshold of minimum length. Precision are increased by keeping as much information on homozygote SNP as possible.

Although changing the threshold in certain criteria set to define ROH did not influence on the detection of ROH in most cases, four criteria need to be commented: (i) First, to account for genotyping errors, the ROH criterion allowed for one heterozygous SNP in a homozygous segment within a window. This criterion created many short false positive ROH, and should be avoided. (ii) Second, by allowing for missing SNP within a window, the detection of ROH was not affected much. Actually, as a SNP dataset became denser, more SNP will be missing because information on some SNP also will be missing. By removing individuals with a call rate less than 0.95 %, it was expected that a maximum of 5 % of the SNP in an individual were missing. Because the amount of ROH on the genome is restricted and proportional to the inbreeding coefficient, the proportion of missing SNP being within a ROH were further reduced. With a limited number of missing SNP per window, it is likely that the number of missing SNP does not affect results much. (iii) Third, maximum average Kb per SNP will on average be positioned less than 5 Kb apart with the HD-panel, implying that the restriction imposed of 50 Kb does not anymore take effect. (iv) Fourth, very few gaps between SNP will be long, especially when low MAF SNP were included and not pruned away, giving small differences in results for the examined gap lengths. Thus, while the need for applying restrictions on the maximum average density per SNP, maximum gap length and number of missing SNP on HD-panel seem redundant, it appears important to keep only homozygous SNP within a window to avoid false positive ROH.

Given that the genotyping error could be controlled by both a GC score threshold [16] and call rate, the remaining low MAF SNP will eventually contribute information to similarity of chromosomal segments passed on from the sire and the dam, i.e. to homozygosity; in support of including this information when determining ROH. Restricting MAF to exceed 0.01 and 0.02 reduced the number of SNP by 14 % and 16 %, respectively, followed by a reduction in the number of ROH detected, mainly short ROH. The data had to pass a genotype quality control, for which the effect of MAF on ROH was examined. Because ROH are continuous homozygote segments dependent on all information available, the method stands out compared to the practice established in GWAS and GS that rely on contrasting effects of genotypes linked up against traits. By removing low MAF SNP in GWAS and GS estimation, it has been succeeded to remove monomorphic SNP that incorrectly were defined as polymorphic and excluded SNP that contribute inaccurately and little to genomic evaluation estimation [17, 18]. Removal of low MAF SNP was also custom in earlier studies within ROH [8, 19, 17, 2, 20], however, recent literature has been in support of including information on low MAF SNP when searching for ROH (Ferenčaković et al, 2013). Thus, because ROH is arranged in continuous segments, it is important to keep as much genomic information as possible, including low MAF SNP, so that ROH will not get split or lost.

By keeping low MAF SNP, an increased amount of short ROH were kept, tails on some stretches were added and gaps were sealed detecting one long ROH instead of two shorter. Because low

MAF SNP often were clustered in long stretches and overrepresented on specific chromosomes, it could indicate either segments of selection signatures or just the fact that some SNP chosen for this chip were not optimal for Norwegian Red. Low MAF SNP have been used to identify selection sweep in cattle [21]. Note that although these SNP are fixed in the population under study, the fact that they are on the HD-panel imply that they still segregates over the populations contributing to the chip. By keeping the low MAF SNP, these SNP will be allowed to be captured in a ROH, mostly by the shortest; that have been exposed to recombination for a long time. Contrary, for more recent selection history, one should look for footprints set out by the longer ROH. Hence, low MAF ROH can signalize selection signatures and trace selection gaining important information on inbreeding.

Conclusions

The detection of ROH was highly influenced by genotyping quality controls, criteria made for identification of ROH and SNP density. A high SNP density improved the estimates of ROH and gained more details. By moving from a low to a high SNP density, several criteria used to define ROH became redundant. We recommend to keep only strictly homozygous segments within a ROH to avoid false positives. Pruning of low MAF SNP are not recommended, as these contributed to loss of information. There is a major need of standards both regarding to genotyping quality controls and to definition criteria when ROH are studied in order to compare results between different studies.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors designed the study, interpreted the findings and revised the manuscript. BH, SAB, and HG prepared the genotype data. BH ran the analysis. BH, JAW, DIV, TM and GK analyzed the results. BH drafted the manuscript. JAW, TM, DIV and GK co-wrote the manuscript.

Acknowledgments

We would like to thank the Norwegian University of Life Sciences for founding this project. We will also acknowledge the breeding organization for dairy cattle in Norway, Geno, by Morten Svendsen and Trygve Roger Solberg for sharing pedigree files and genotyping data. At last we want to thank Professor Johann Sölkner from the University of Natural Resources and Life Sciences (BOKU) for welcoming Borghild Hillestad to his group and expanding her knowledge on ROH.

References

1. Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. Am J Hum Genet. 1999;65(6):1493-500. doi:10.1086/302661.

2. Kirin M, McQuillan R, Franklin C, Campbell H, McKeigue P, Wilson J. Genomic runs of homozygosity record population history and consanguinity. PLoS One. 2010;5(11):e13996.

3. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research. 2003;13(4):635-43. doi:10.1101/gr.387103.

4. McQuillan R, Leutenegger A, Abdel-Rahman R, Franklin C, Pericic M, Barac-Lauc L et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83(3):359 - 72.

5. Bolormaa S, Pryce JE, Hayes BJ, Goddard ME. Multivariate analysis of a genome-wide association study in dairy cattle. Journal of Dairy Science. 2010;93(8):3818-33. doi:10.3168/jds.2009-2980.

6. Nishimura S, Watanabe T, Mizoshita K, Tatsuda K, Fujita T, Watanabe N et al. Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. Bmc Genetics. 2012;13. doi:10.1186/1471-2156-13-40.

7. Kim ES, Cole JB, Huson H, Wiggans GR, Van Tassell CP, Crooker BA et al. Effect of Artificial Selection on Runs of Homozygosity in US Holstein Cattle. Plos One. 2013;8(11). doi:10.1371/journal.pone.0080813.

8. Ferenčaković M, Hamzić E, Gredler B, Solberg TR, Klemetsdal G, Curik I et al. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics. 2012:n/a-n/a. doi:10.1111/jbg.12012.

9. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819-29.

10. Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD et al. Distribution and location of genetic effects for dairy traits. Journal of Dairy Science. 2009;92(6):2931-46. doi:10.3168/jds.2008-1762.

11. Keller M, Visscher P, Goddard M. Quantification of inbreeding due to distant ancestors and its detection using dense SNP data. Genetics. 2011.

12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 812007.

13. Ferenčaković M, Sölkner J, Curik I. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics, selection, evolution : GSE. 2013;45(1):42-. doi:10.1186/1297-9686-45-42.

14. Rodríguez-Ramilo ST, Fernández J, Toro MA, Hernández D, Villanueva B, editors. Genome-wide estimates of effective population size in the Spanish Holstein population. WCGALP; 2014; Vancouver, Canada.

15. Gómez-Romano F, Sölkner J, Villanueva B, Mézáros G, Cara MARd, O'Brien AMP et al., editors. Genomic estimates of inbreeding and coancestry in Austrian Brown Swiss cattle. WCGALP; 2014; Vancouver, Canada.

16. Illumina. Illumina GenCall Data Analysis Software. www.illumina.com. 2005. http://res.illumina.com/documents/products/technotes/technote_gencall_data_analysis_software.pdf.

17. Edriss V, Guldbrandtsen B, Lund MS, Su G. Effect of marker-data editing on the accuracy of genomic prediction. Journal of Animal Breeding and Genetics. 2013;130(2):128-35. doi:10.1111/j.1439-0388.2012.01015.x.

18. Wiggans GR, Sonstegard TS, VanRaden PM, Matukumalli LK, Schnabel RD, Taylor JF et al. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. Journal of Dairy Science. 2009;92(7):3431-6. doi:http://dx.doi.org/10.3168/jds.2008-1758.

19. Howrigan D, Simonson M, Keller M. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. BMC Genomics. 2011;12(1):460.

20. Silió L, Rodríguez MC, Fernández A, Barragán C, Benítez R, Óvilo C et al. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. Journal of Animal Breeding and Genetics. 2013:n/a-n/a. doi:10.1111/jbg.12031.

21. Ramey HR, Decker JE, McKay SD, Rolf MM, Schnabel RD, Taylor JF. Detection of selective sweeps in cattle using genome-wide SNP data. Bmc Genomics. 2013;14. doi:10.1186/1471-2164-14-382.

Table 1: Genotyping quality controls

Genotyping quality controls done on the Illumina HD-panel for 384 bulls in Norwegian Red.

Genotyping quality control	Remaining SNP	Lost # SNP	Lost in percent
Initial dataset	777,962	0	0
Autosomal SNP only	735,293	42,669	5.48
Animals with > 95% call rate	735,293	0	0
SNP with > 90% call rate	708,620	26,673	3.63
Hardy Weinberg Equilibrium (p <1e-06)	707,609	1,011	0.14
SNP with MAF< 0.01	610,885	96,724	13.67
SNP with MAF< 0.02	597,454	13,431	2.20

Table 2: Datasets used to detect ROH

An overview over different SNP-datasets used to find ROH in 381 Norwegian Red bulls.

Density	Exact # of SNP	SNP pr Kb
Main density sets		
53K	53,129	0.0177
71K	70,839	0.0236
94K	94,452	0.0315
126K	125,937	0.0420
168K	167,917	0.0560
224K	223,890	0.0746
299K	298,521	0.0995
398K	398,029	0.1327
531K	530,706	0.1769
708K	707,609	0.2359
MAF sets		
$597K_{MAF}$	597,454	0.1992
$611K_{MAF}$	610,885	0.2036

Table 3: Constraints set to detect ROH in Norwegian Red

SNP density	SNP pr window (5,000 Kb)	Min. # homozygous SNP	Min. # homozygous Kb	# hetrozygote SNP allowed pr window	# missing SNP allowed pr window	Max. gap length (Kb)	Max. avg. Kb pr SNP
			Main density sets				
53K	88.5	15	2,000	0	1	1,000	150
71K	118.1	15	2,000	0	1	1,000	150
94K	157.4	15	2,000	0	1	1,000	150
126K	209.9	25	1,000	0	2	500	150
168K	279.9	25	1,000	1	2	500	150
224K	373.2	25	1,000	1	2	250	50
299K	497.5	25	1,000	1	2	250	50
398K	663.4	50	500	1	3	250	50
531K	884.5	50	500	1	3	250	50
708K	1,179.3	50	500	1	3	250	50
			Variants of HD-panel				
$708\text{K}_{\text{Alt1}}$	1,179.3	50	500	0	3	250	50
$708\text{K}_{\text{Alt2}}$	1,179.3	15	2,000	0	1	1,000	150
$708\text{K}_{\text{Alt3}}$	1,179.3	25	1,000	0	2	500	150
$708\text{K}_{\text{Alt4}}$	1,179.3	25	1,000	1	2	250	50
$708\text{K}_{\text{Alt5}}$	1,179.3	50	500	0	1	250	50
$708\text{K}_{\text{Alt6}}$	1,179.3	50	500	0	3	1,000	50
$708\text{K}_{\text{Alt7}}$	1,179.3	50	500	0	3	250	150
$708\text{K}_{\text{Alt8}}$	1,179.3	50	500	0	15	250	50
$708\text{K}_{\text{Alt9}}$	1,179.3	50	500	0	3	68	50
$708\text{K}_{\text{Alt10}}$	1,179.3	50	500	0	15	68	50

			MAF sets				
$597K_{MAF}$	995.8	50	500	1	3	250	50
$611K_{MAF}$	1,018.1	50	500	1	3	250	50

Table 4: Detected ROH

Average number of ROH detected per individual, grouped into lengths of the segment in 381 Norwegian Red.

								Total
SNP density	0.5-1Mb	1-2Mb	2-4Mb	4-8Mb	8-16Mb	>16Mb	Total	>2Mb
				Main density sets				
53K	-	-	9.8	8.0	4.0	1.4	23.2	23.2
71K	-	-	12.9	8.0	3.9	1.4	26.2	26.2
94K	-	-	13.1	8.0	3.9	1.4	29.4	29.4
126K	-	22.1	13.1	8.0	3.9	1.3	48.4	26.7
168K	-	36.2	14.0	8.0	3.9	1.5	63.6	27.4
224K	-	33.1	13.5	8.2	3.9	1.4	60.1	27.0
299K	-	30.4	13.6	8.2	3.9	1.3	57.4	27.0
398K	153.8	28.6	13.4	8.1	3.9	1.3	209.1	26.7
531K	142.4	27.4	13.4	8.0	3.9	1.3	196.4	26.6
708K	131.1	26.3	13.4	8.1	3.9	1.3	184.1	26.7
				Variants of the HD-panel				
$708K_{Alt1}$	89.3	23.0	14.1	8.4	3.6	1.0	139.4	27.1
$708K_{Alt2}$	-	-	14.4	8.2	3.5	0.9	27.0	27.0
$708K_{Alt3}$	-	23.2	14.0	8.3	3.7	1.0	50.2	27.0
$708K_{Alt4}$	-	26.5	13.5	8.1	3.8	1.3	53.2	26.7
$708K_{Alt5}$	90.0	24.0	14.6	8.3	3.4	0.9	141.2	27.2

708K$_{\text{Alt6}}$	89.4	23.2	13.9	8.3	3.7	1.1	139.5	27.0
708K$_{\text{Alt7}}$	89.3	23.0	14.1	8.4	3.6	1.0	139.4	27.1
708K$_{\text{Alt8}}$	89.3	23.0	14.1	8.3	3.6	1.0	139.3	27.0
708K$_{\text{Alt9}}$	89.1	24.1	14.8	8.6	3.3	0.7	140.6	27.4
708K$_{\text{Alt10}}$	89.1	24.0	14.8	8.6	3.3	0.7	140.5	27.4
MAF sets								
597K$_{\text{MAF}}$	120.3	25.3	13.0	8.0	3.8	1.3	171.7	26.1
611K$_{\text{MAF}}$	121.9	25.5	13.0	8.0	3.8	1.3	173.5	26.1

Table 5: Chromosome wise loss of SNP by removing Low MAF SNP

Total loss of SNP per chromosome and short ROH (0.5-1Mb) by pruning for low MAF SNP and average heterozygosity (Het) in 381 Norwegian Red genotyped with an Illumina HD-panel.

BTA	Size of BTA in Mb *	Total SNP	No ROH (0.5-1 Mb)	MAF<0.01		MAF<0.02		
				% SNP	% ROH	% SNP	% ROH	Het
1	158	45,007	10.9	13.9	5.6	16.2	5.9	0.351
2	137	38,738	9.0	14.6	4.2	16.5	5.4	0.358
3	121	34,229	7.7	12.7	5.7	15.5	6.9	0.355
4	121	33,749	5.7	13.1	4.2	15.2	4.3	0.354
5	121	33,394	7.3	15.2	6.8	17.7	7.8	0.346
6	119	34,441	5.5	11.9	4.3	13.9	4.6	0.353
7	113	31,831	6.1	14.8	10.8	16.9	13.3	0.365
8	113	32,423	7.0	28.7	9.2	30.8	11.4	0.349
9	106	29,999	5.9	14.0	5.4	16.3	5.4	0.353
10	104	29,350	4.9	11.0	8.4	13.0	8.9	0.357
11	107	30,949	5.9	10.5	3.1	12.9	3.9	0.358
12	91	25,011	4.0	12.7	5.3	15.1	5.9	0.360
13	84	22,704	5.2	23.9	16.8	27.0	18.6	0.343
14	85	23,972	5.4	25.4	16.9	28.3	19.7	0.341
15	85	23,509	4.7	11.1	5.2	13.6	6.8	0.352
16	82	23,222	5.0	12.5	8.1	14.6	8.7	0.360
17	75	21,417	3.2	9.8	7.1	12.4	7.8	0.354
18	66	18,443	3.0	8.2	12.6	10.2	13.6	0.360
19	64	18,047	2.9	8.5	5.1	11.4	12.7	0.355
20	72	20,801	3.4	8.5	9.3	10.6	10.4	0.359

21	72	20,296	4.1	12.9	6.6	14.9	9.3	0.352
22	61	17,356	2.7	7.4	1.3	9.9	1.5	0.357
23	53	14,499	1.1	9.8	1.7	11.8	0.7	0.358
24	63	18,030	3.1	13.0	7.8	14.8	10.5	0.362
25	43	12,358	1.0	7.2	0.5	9.3	1.1	0.364
26	52	14,707	1.8	8.0	9.6	10.6	9.9	0.348
27	45	12,690	1.3	7.8	1.8	10.3	2.3	0.351
28	46	12,456	1.5	7.7	1.9	9.2	2.6	0.366
29	52	13,981	1.9	9.1	3.7	11.1	4.5	0.351
Total	2,511	707,609	131.1	13.4	7.0	15.7	8.3	0.355

* (http://www.ncbi.nlm.nih.gov/genome?term=bos%20taurus)

Figure 1: Cumulative frequency of ROH detected in Norwegian Red

Cumulative frequency of the number of detected ROH by length of ROH ranging between minimum 0.5 to maximum 58.7 Mb in 381 Norwegian Red genotyped with an Illumina HD-panel ($708K_{Alt1}$).

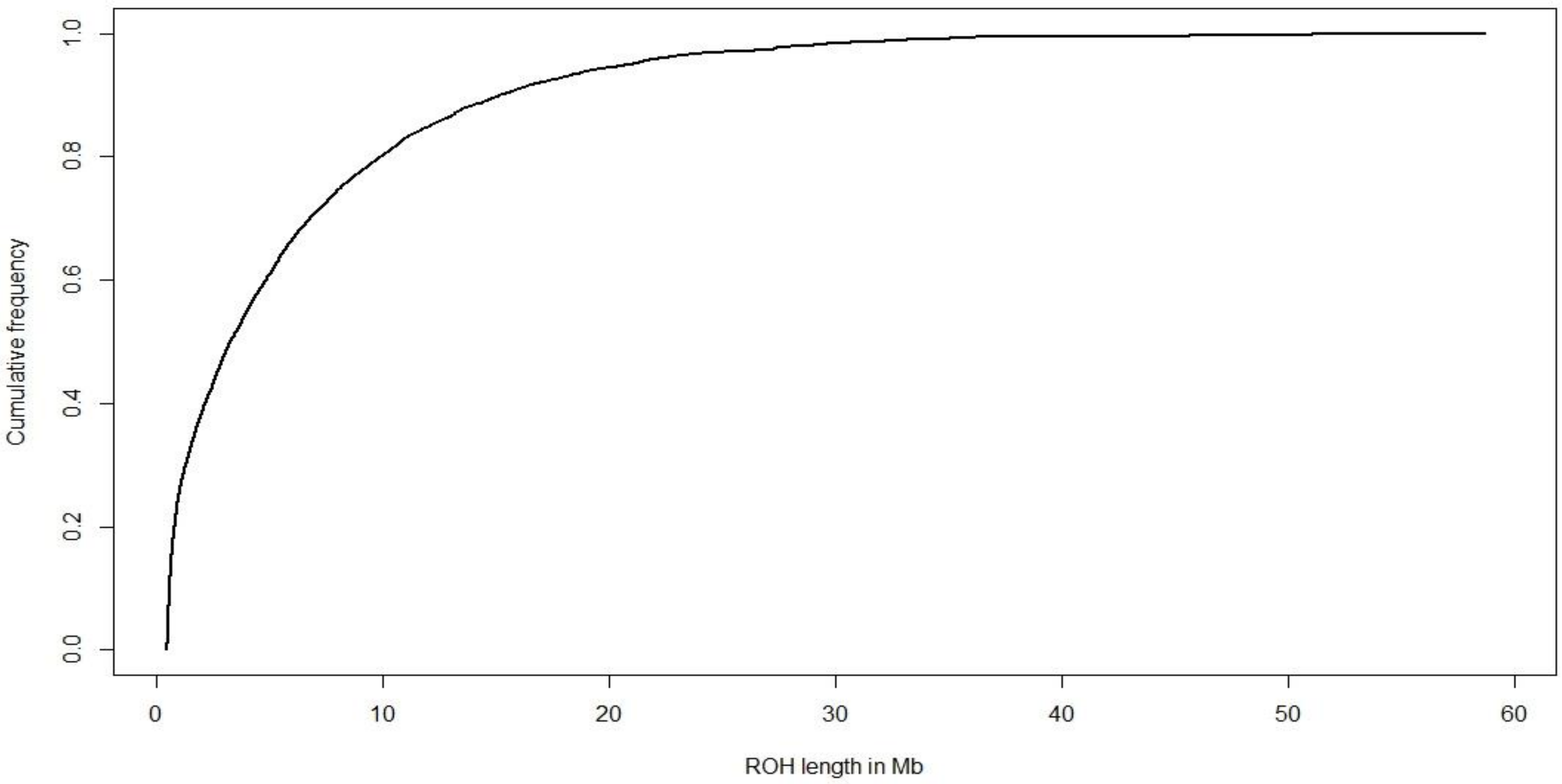

Cumulative frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
ROH length in Mb

Figure 1

Paper II

Estimating rate of inbreeding and effective population size using genomic data in Norwegian Red

Borghild Hillestad, John Arthur Woolliams, Theo Meuwissen, Dag Inge Våge,
Gunnar Klemetsdal

Submitted to Genetic Selection Evolution

Estimating rate of inbreeding and effective population size using genomic data in Norwegian Red

Borghild Hillestad[1], John Arthur Woolliams[1,2], Theo Meuwissen[1], Dag Inge Våge[1,3], Gunnar Klemetsdal[1]

[1]Department of Animal and Aquacultural Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, N-1432 Ås, Norway

[2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

[3]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, N-1432 Ås, Norway

Borghild Hillestad borghildhillestad@gmail.com

John Arthur Woolliams john.woolliams@roslin.ed.ac.uk

Theo Meuwissen theo.meuwissen@nmbu.no

Dag Inge Våge daginge.vage@nmbu.no

Gunnar Klemetsdal: gunnar.klemetsdal@nmbu.no

Corresponding author: Borghild Hillestad

Abstract

Background: Traditionally, rate of inbreeding and effective population sizes have been estimated by use of pedigree data. The objective of this study was to compare ΔF and Ne from runs of homozygosity, observed homozygosity and pedigree and for genetic measures to find the effect of SNP density, genotyping quality controls and imputation.

Methods: Inbreeding coefficients (F) were estimated by utilizing genomic data, both by runs of homozygosity (ROH) and by observed homozygosity. These two genomic inbreeding measures and a traditional inbreeding coefficients from pedigree was in a ln(1-F) format, regressed on either (i) year of birth or (ii) complete generation equivalent (CGE) to estimate the rate of inbreeding (ΔF) and effective population size (Ne). Two dataset were exploited: (i) 384 Norwegian Red bulls genotyped with the Illumina HD-panel containing 777K SNP-markers, and (ii) 3,289 Norwegian Red bulls genotyped with a 54K Illumina BeadChip and/or 25K Affymetrix, with imputations done both ways if needed. The pedigree of these two datasets extended as far back as 1875.

Results: The pedigree suffered of a threshold effect, and was found too young to give an asymptotic estimate of ΔF and Ne alone, and should rather be based on genomic measures regressed on CGE. From observed homozygosity, a Ne of 57.5 animals was obtained, approximately 1/3 of what was obtained by ln(1-F_{Ped}) regressed on year of birth.

Conclusions: Observed homozygosity gave more stable results, accounting for more homozygosity than ROH. By regressing inbreeding coefficients on CGE a better fit by a higher R^2 was achieved, compared to year of birth. Further, it was recommended to keep all low MAF SNP in analysis.

Keywords: Runs of Homozygosity (ROH), Rate of Inbreeding (ΔF), Genomic Inbreeding, Observed Homozygosity, Effective Population Size (Ne), Cattle

Background

In commercial livestock breeds, the inbreeding coefficient (**F_{Ped}**) of an individual is typically estimated based on the pedigree [1]. The individual inbreeding coefficient is the probability of identity by descent of a selection free neutral allele relative to that of the base population, with 2N different alleles. With pedigree errors, contemporary individuals may have different depths of pedigree available, affecting not only $F_{Ped,}$ but also the rate of inbreeding (**ΔF**) and the effective population size (**Ne**) estimates. A genome based inbreeding coefficient has the potential to circumvent these problems, and would be particularly useful for assessing Ne in livestock populations lacking a complete herdbook, or in wild populations.

Methods to estimate Ne using genomic data have been developed using linkage disequilibrium (**LD**); such as chromosomal segment homosygosity and r-squared, but there are indications that these methods are weak in addressing the most recent generations [2-5]. To address the latter, Saura et al. [6] recently compared estimation of ΔF and Ne in Iberian pigs from pedigree and genomic data. Inbreeding rates were obtained by regressing the natural logarithm of (1-F) on year of birth, where individual F was estimated either from genealogical or molecular coancestry. Observed homozygosity has also been used by Bjelland et al. [7] and Silió et al. [8] to measure genomic inbreeding.

Alternatively, the individual inbreeding coefficient (**F_{ROH}**) can be calculated from runs of homozygosity (**ROH**); stretches of homozygous segments present in the genome caused by parents transmitting identical haplotypes to their offspring [9]. By looking at the ratio between the total length of ROH in an individual and the length of the genome, an observed inbreeding coefficient (**F_{ROH}**) can be calculated [10]. Broman and Weber used molecular markers to demonstrate the relationship between the length of the homozygous segment and the length in time from the common ancestor in a human dataset. Homozygous segments originating from a more recent ancestor are expected to be longer than segments from an ancient ancestor due to the increasing number of recombination events over time [2].

Observed homozygosity has proven to give a parameter with high correlations to both pedigree and ROH based estimates, but differs from ROH by identifying all homozygosity instead of clustered homozygosity [8, 7]. The strength of ROH is claimed to be that it extracts SNP that are identical by decent (**IBD**) from markers that are only IBS, arising from more recent inbreeding. Therefore, ROH may be more suited for estimating more recent Ne. One weakness of ROH is the ambiguity of definition, which has previously been addressed by Hillestad et al. [11].

This study carried out using genomic and pedigree data from Norwegian Red. With a well-documented herdbook and high density genotyping data available over time, this breed qualifies as a good test population for comparing genomic and pedigree based inbreeding parameters. The objective was to compare ΔF and Ne based on genomic data from ROH, observed homozygosity and pedigree either in combination or separately, and to investigate the effects of SNP density, minimum lengths to detect ROH, genotyping quality controls and imputation.

Material and Methods

Population and pedigree data

This study was based on a total of 2,372 Norwegian Red bulls born between 1975 and 2009. Both genotype and pedigree data were available for all animals, although the amount of genotype data varied between subgroups of animals. In total, 1,116 bulls were genotyped with the 54K Illumina BeadChip [12], and 1,704 bulls had been genotyped with the 25K Affymetrix chip [13]. A total of 448 bulls were genotyped with both the 25K and the 54K chips, while those genotyped by only one of the chips were imputed using Beagle [14]. A subgroup of 375 bulls had also been genotyped with the 777K Illumina HD-panel [15].

The pedigree data of this population extended as far back as 1875. The pedigree depth was summarized by the complete generation equivalent (**CGE**) using Pedig [16] also estimated by the equation of Maccluer et al. [17]:

$$\mathrm{CGE} = \frac{1}{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} \sum_{i=1}^{n_j} \frac{1}{2^{g_{ij}}} \qquad (1)$$

Here *N* refers to number of genotyped animals; $n_{j,}$ the total number of ancestor of animal *j* in the population under this study; and g_{ij}, the number of generations between *j* and its ancestor *i*. The CGE were traced back no more than 20 generations per each individual due to limitations in Pedig.

Individual inbreeding coefficients were calculated using RelaX2 [18], which uses the algorithm of Meuwissen & Luo [19]. Inbreeding coefficients (**F_{Ped}**) were derived from the pedigrees where the base population was considered to be those with unknown parents in the historical records, ignoring their depth of pedigree.

Quality control and SNP density of genotype data

Two methods of quality controls were used in this study: Industry quality controls (**IQ**) and high density quality controls (**HDQ**).

IQ were based upon the 54K data of the full set of 2,372 animals including imputed genotypes (Table 1). As this group had been targeted towards GS and the calculation of GEBV, the following genotyping quality controls had already been carried out: (i) removal of animals with an individual call rate < 97 %, (ii) deletion of Mendelian errors for animals with known parents, (iii) removal of SNP with Mendelian error rate > 2.5 %, (iv) deletion of SNP with a call rate < 25 %, and (v) removal of SNP with MAF < 0.05. After these criteria had been applied, a dataset of

48,249 SNP remained (**48K$_{\mathbf{GS}}$**). The IQ was also applied to the 375 bulls genotyped with the HD-panel resulting in a density of 539,665 SNP (**540K$_{\mathbf{GS}}$**).

A further quality control was performed for the 375 bulls genotyped with the HD-panel (HDQ). This was done to optimize the genotypes for estimating ROH, and the conditions were as follows: (i) exclusion of markers on sex-linked chromosomes, (ii) minimum call rate per SNP > 90 %, (iii) deviation from Hardy-Weinberg ($P > 10^{-6}$), and (iv) genotypes for fewer than 95 % of markers. After this a total of 707,609 SNP remained (**708K**), and 3 animals were removed because of failing criteria iv (Table 1).

To generate different SNP densities from the HD-panel, the 708K-set was sequentially pruned to give nine less dense subsets. The first pruning removed every fourth SNP, by physical order, from the 708K set to obtain a subset of 530,706 SNP (**531K**). This procedure was repeated by removing every fourth SNP from the 531K-set, to obtain a **398K** set, and a further seven times to give the smallest subset (**53K**). All densities and subsets are shown in Table 1.

Derivation of inbreeding coefficients from genomic data

ROH were identified with PLINK 1.07 [20] for each animal. PLINK operates with sliding window, analyzing a segment of 5 Mb at a time. The identifications of ROH required specifications of criteria, and values used were based on the conclusions of Hillestad et al. [11]. For criteria, (i) the minimum length of a ROH was either 0.5 or 2 Mb, (ii) no heterozygote SNP was allowed within a ROH and (iii) Minimum numbers of SNP in a ROH were set to the expected number of SNP in a 500 Kb segment at the given density. All other criteria depended on the density of the SNP panel as shown in Table 2.

Individual inbreeding coefficients from ROH were calculated as followed;

$$F_{ROH} = \frac{\sum L_{ROH}}{\sum L_{AUTO}} \qquad (2)$$

where $\sum L_{ROH}$ is an individual's total ROH length, and $\sum L_{AUTO}$ is its total length of autosome covered by SNP which was 2.51 Gb [10]. This coverage represent 83.67 % of the total autosomal genome. A further individual inbreeding coefficient (**F_{Hom}**) was estimated on observed fraction homozygous SNP for each individual ignoring haplotypes:

$$F_{Hom} = O(Hom) / N(NM) \qquad (3)$$

where N(NM) was defined as the number of non-missing genotypes and O(Hom) the amount of observed homozygosity.

Expected relationship of genomic and pedigree F-values

F_{ROH} and F_{Hom} are values based on observed homozyosity, while the F_{Ped} will be a measure of expected homozygosity and will depend upon where the base population is set. A relationship of the form:

$$\left(1-\mathrm{F}_y\right)=\left(1-\mathrm{F}_{\mathrm{Ped}}\right)\left(1-\mathrm{F}_{\mathrm{Pop}}\right) \tag{4}$$

might be anticipated, where *y* refers to ROH or observed homozygosity. $\mathrm{F}_{\mathrm{Pop}}$ is common to all individuals in the population [21]. Taking the logarithm to linearize gave:

$$ln\left(1-\mathrm{F}_{\mathrm{y}}\right)=ln\left(1-\mathrm{F}_{\mathrm{Ped}}\right)+ln(1-\mathrm{F}_{\mathrm{Pop}}) \tag{5}$$

Then the following regression model applied on an individual basis (*i*):

$$ln\left(1-\mathrm{F}_{\mathrm{y}}\right)_i=y_i=\mu+\beta\cdot ln\left(1-\mathrm{F}_{\mathrm{Ped}}\right)_i+\mathrm{e}_i \tag{6}$$

where μ is a constant expected to equal $ln(1\text{-}\mathrm{F}_{\mathrm{Pop}})$. To test the regression the following null hypothesis were set: H_0: $\beta = 1$ against the alternative H_1: $\beta \neq 1$

Inbreeding rate and effective population size

By utilizing theory from inbreeding of the idealized population and CGE from analysis of pedigree data, the following equation was set [21, 17]:

$$\left(1-\mathrm{F}_{\mathrm{y}}\right)=\left(1-\Delta\mathrm{F}_{\mathrm{y}}\right)^t(1-\mathrm{F}_{\mathrm{Pop}}) \tag{7}$$

where *y* referred to pedigree, ROH or observed homozygosity and *t* referred to CGE. To make this linear, the natural logarithm was taken, leading to:

$$ln\left(1-\mathrm{F}_{y}\right)=t ln\left(1-\Delta\mathrm{F}_{\mathrm{y}}\right)+ln(1-\mathrm{F}_{\mathrm{Pop}}) \qquad (8)$$

which was individually fitted by the following linear regression equation:

$$ln\left(1-\mathrm{F}_{\mathrm{y}}\right)_{i}=y_{i}=\mu+\beta t_{i}+\mathrm{e}_{i} \qquad (9)$$

where μ is $ln(1\text{-}\mathrm{F}_{\mathrm{Pop}})$ from [4] and β is the regression coefficient of CGE on *y*. Estimates of ΔF and Ne was obtained by the following equations:

$$\begin{aligned}\Delta\mathrm{F}&=1-\mathrm{e}^{\beta}\\ \mathrm{Ne}&=(2\Delta\mathrm{F})^{-1}\end{aligned} \qquad (10)$$

Correspondingly, one can regress on year of birth rather than on CGE, and then estimate ΔF by multiplying by the generation interval (**L**):

$$\Delta\mathrm{F}=(1-\mathrm{e}^{\beta})\mathrm{L} \qquad (11)$$

and eventually estimating Ne with formula [10]. L was obtained by regressing CGE on year, resulting in 5 years per generation (Figure 1).

Comparisons made in study

The three measures of individual inbreeding (F_{Ped}, F_{ROH} and F_{Hom}) and the two values of Ne (either by regressing on CGE or year of birth) obtained from each of these measures were compared for different genomic approaches. The effect of SNP density ranging from 53K to 708K was examined using the panel obtained from pruning the 375 animal with HD genotypes using HDQ. The effect of minimum length was examined by comparisons of results from 53K and 708K using the HDQ, with minimum lengths of 0.5 Mb and 2 Mb, respectively. The effect of the approach to quality control was examined by comparisons of results from $48K_{GS}$ and $540K_{GS}$ using IQ, with 53K and 531K using HDQ. The effect of imputation was examined by comparing the results using $48K_{GS}$ panel with the 448 being operating with real genotypes with the 1,704 and 1,116 animals that had been imputed.

Results

By plotting ln(1-F_{Ped}) against ln(1-F_{ROH}) and ln(1-F_{Hom}), it was obvious that the pedigree suffered of a threshold effect, and needed yet a greater depth to reach a steady state (Figure 1). Even though the genotypes showed huge differences between animals in the genomic data, the values of ln(1-F_{Ped}) did not seem able to present that difference, and all except highly inbred individuals were placed at the upper corner. This gave inbred animals too much weight to the regression. Values from ln(1-F_{ROH}) and ln(1-F_{Hom}) showed a nice distribution to both CGE and year for birth where the smoothing line followed the regression line well. Plots of ln(1-F_{ROH}) and ln(1-F_{Hom}) against different SNP densities from 53 to 708K showed how a higher density reduced errors (data not shown). By regressing ln(1-F_y) on CGE, the R^2 of the regression was doubled relative to when year of birth was used as the explanatory variable (Table 3). ROH gave the lowest R^2, mostly decreasing with lower density. Pedigree regressed on CGE was observed with the highest R^2 of 0.13, but according to Figure 1, it did not give the best estimate of inbreeding. The best fit when measuring inbreeding was therefore ln(1-F_{Hom}) using a 708K density regressed on CGE, providing a R^2 of 0.12.

The effect of SNP density

Average F_{ROH} had a tendency to increase with increased density from 53K to 708K (Table 4). This was accompanied by a small increased correlation between F_{ROH} and F_{Hom}. Apart from this correlation, F_{Hom} did not seem to be affected by SNP-density. The slopes of the regressions of ln(1-F_{ROH}) and ln(1-F_{Hom}) on ln(1-F_{Ped}) show values slightly larger than 1 for all SNP densities with no particular trend (Table 5). Molecular F-values show slight, but not significantly different from the pedigree estimate, and observed homozygosity consequently provided higher values than ROH. In general the slopes of these regressions always ended up higher than 1 in all HDQ-sets, irrespective of SNP density. ΔF_{ROH} increases and Ne decreases with density (Table 3). In contrast, observed homozygosity gave larger estimates of ΔF, but did not seem to increase with density. Both molecular ΔFs were greater than when predicted by pedigree. All estimates of ΔF were lower when estimated by year of birth than by CGE. By year of birth, the estimate had a bigger variation in Ne between the highest and lowest density compared to estimates based on CGE. In summary, molecular, and to some degree high density for ROH, seemed to increase the rate of inbreeding compared to pedigree estimates, resulting in lower molecular Ne compared to pedigree Ne.

The effect of minimum length

When restricting ROH to 2 Mb, a higher density did not increase average F_{ROH}, that was stabilized at the 53K level (Table 4). Neither was the correlations to pedigree much affected by the restrictions. Although the slopes of the regression of ln(1-F_{ROH}) on ln(1-F_{Ped}) was somewhat reduced, it was still not significantly different from zero (Table 5). For increased minimum length, ΔF was not much affected relative to that obtained at 53K with a minimum length of 0.5 Mb; both by year of birth and by CGE (Table 6).

The effect of genotyping quality control

IQ tended to give lower average Fs than HDQ, where ROH gave larger differences than observed homozygosity (Table 4). F_{ROH} also contributed to a slightly higher correlation to F_{Ped} in IQ than in HDQ. Genotyping quality control had a considerably effect on the regression of molecular Fs on pedigree (Table 5). When values from HDQ in general were entirely consistent to 1 or had a tendency of being greater than 1, IQ was interfering, especially with ROH; $540K_{GS}$ was extremely affected, and gave a slope as low as 0.48, reflecting only 48 % of the total variation in F_{Ped}-values. With IQ, both ROH and observed homozygosity gave approximately twice as low ΔF compared to HDQ (Tables 6 and 3). This had a big effect on Ne contributing to an impression of a high Ne, especially when ΔF was regressed on year of birth. With IQ, Ne was highly raised both by regressing on year of birth and by CGE compared to HDQ. Thus, genotyping quality control seemed to have a great influence on all ΔF estimates from ROH, but also an effect on observed homozygosity.

The effect of imputation

Imputation of genotypes did not seem to affect molecular Fs, and their correlations to either each other or to pedigree (Table 7). But when studying the relationship between molecular Fs and F_{Ped}, imputation from Beagle leaded to a further interference between pedigree and genomic F (Table 8). Although the Both-set (containing both 25K and 54K without imputation) only had a slope of 0.92 for F_{Hom}, not being able to explain all the variation in F_{Ped}, and 0.83 using F_{ROH} due to IQ, both the 25K and the 54K sets revealed a further noise of the amount of variation being caused by imputation in Beagle. Table 9 and the regressions done on ln(1-F_{Ped}) illustrated that the animals of the 54K set had a slightly higher ΔF than the other two sets, which reflected the genomic results as well. According to the findings where ln(1-F_{Hom}) regressed on CGE gained the best R^2 and the best fit of the regressions, it was notable that the Both-set gave more stable ΔF than the other two imputed groups when comparing them to ln(1-F_{Ped}) regressed on CGE.

Discussion

The goal of this study was to compare inbreeding ΔF and Ne based on genomic data with the corresponding ΔF and Ne from pedigree. The study showed how F_{Ped} underestimated ΔF compared to molecular F, because the pedigree was not deep enough. It also demonstrated how only F_{ROH} was sensitive to SNP density, while both F_{ROH} and F_{Hom} were affected by genotyping quality controls, mainly pruning for low MAF, and imputation from Beagle.

Pedigree appeared to be influenced by a threshold effect, implicating that a pedigree needed to reach a certain amount of generations before it stabilized F. Therefore, a considerable spread in marker based inbreeding was observed for small values of pedigree inbreeding. In this pedigree, on average 7-8 generations was recorded, and did not seem to be deep enough. That way, the animals with the highest F_{Ped} were credited with most weight in the regressions. Thus, pedigree inbreeding contained less information than the corresponding measures from markers, demonstrated by the threshold effect. In consequence, the rate of inbreeding from pedigree gave lower estimates.

Increased marker density was of great importance to the average level of F_{ROH}, but did not have the same effect on F_{Hom}. With reduced density, SNP were still evenly distributed across the genome and random due to the total amount of homozygosity, but not random to clustered homozygosity. Thus, because observed homozygosity had less assumptions compared to ROH, and did not rule any homozygosity out, this approach gave more stable and consistent estimates across SNP densities. Despite this, increased density resulted in a slightly better fit for $\ln(1-F_{Hom})$ than reduced density, implying that individual F_{Hom} was more precisely determined by a high SNP density.

Due to assumptions for ROH; by raising the threshold for minimum length to define ROH, even more information was removed and the estimates from high densities were set back at a lower density level. Thus, by adding more constraints to ROH, the distance between the results from ROH and observed homozygosity was increased and the estimates from ROH were aggravated.

Too many constraints may be the reason why regressions of ln(1-F_{Hom}) gave a higher R^2 than ln(1-F_{ROH}). In consequence, ∆F increased with increased SNP density for ROH, but not for observed homozygosity.

By considering R^2-values of the regressions, CGE was found to be a better explanatory variable than year of birth in this population. CGE relied on the pedigree, and was easily obtained in a population where its genealogy was recorded. In the wild, however, one would need to regress on time, and sample data over a relevant time span, taking the generation length into account. Also, in populations where CGE has no variation, for instance for some populations in the fish industry, the parameter would not have the same effect as in the Norwegian Red population.

When low MAF SNP were removed, the slope of the regression of molecular F on F_{Ped} was consistently reduced as well as ∆F (Table 3). Low MAF SNP may result from genotyping error where monomorphic SNP falsely detects variation in a few animals, but they can also result from random genetic drift, recent mutation and selection resulting in near complete fixation [22]. ROH are continuous, homozygote stretches, where low MAF SNP contributes information to similarity of the homozygous stretches that may have been passed on from the parents. Slopes significantly lower than 1 by regressing F_{ROH} on F_{Ped} have also been observed in other studies. Recently, Rodríguez-Ramilo et al. [23] found a slope of 0.79 when F_{ROH} was regressed on F_{Ped} using a 37K density in Spanish Holstein. Similarly, Gómez-Romano et al. [24] obtained a slope of 0.71 in Austrian Brown Swiss. While Rodríguez-Ramilo et al. [23] used a minimum length for ROH of 1 Mb, Gómez-Romano et al. [24] used 4 Mb. Both studies allowed 1 heterozygote

SNP within a run, which may have contributed to false positive ROH, especially for low SNP densities [11]. In addition to low SNP density, neither of these articles mentioned how low MAF SNP were handled, questioning whether this also may have contributed to the reduced slope. Removal of low MAF SNP will split and shorten ROH, because these SNP are often clustered together or attached to a ROH. Therefore, pruning of low MAF SNP will remove important inbreeding information. In general, correct genotyping quality controls and ROH constraints are vital to get truthful estimates, because small adjustments on ΔF will change Ne dramatically. Misaligned preparations of the genotypes may even give the impression of a higher Ne than predicted by pedigree as shown by the IQ sets, which is why genotyping quality controls need to be customized ROH and the constraints on ROH carefully considered.

In the IQ sets, all SNP with MAF < 0.05 were removed for all individuals, regardless of the allele frequency of the SNP in the founder population. The SNP were not selected for their initial MAF but for their 'population-wide MAF > 0.05', which may be closer to the current MAF of the SNP than the initial MAF (since most of the genotyped animals were currently alive bulls). This could be an explanation on why β moved below 1 when genomic F was regressed on F_{Ped} (Table 8). Consider a set of SNP with initial MAF = 0.05: Most of these SNP would be expected to drift to a MAF below 0.05, but if this happened their population-wide MAF would be below 0.05, and excluded by IQ. Only SNP who happened to drift to higher MAF than 0.05 would be included by IQ, and their heterozygosity would be increased. Hence, the selection of the SNP from IQ favored SNP that either had drifted to a high frequency or had a high heterozygosity. The latter may have resulted in the bias indicated by the β-values < 1.

The relationship between ln(1-F) from genomic data and F_{Ped} was disturbed by imputation from Beagle, which relies on linkage disequilibrium without utilizing known relationships [14]. This could be an element that causes error. By making use of pedigree information as well, it would be possible to compare alleles within family [25]. In this way, pedigree would operate as an extra quality check of the imputation. Imputation of genotypes from two different chips is an cost-effective method to gain more information to many animals based on a small reference population [26], and it would be preferable to utilize imputed data to estimate inbreeding. In order to impute SNP genotypes, it is custom to remove SNP with MAF < 0.05, which may be a problem to inbreeding measurements, and in addition to a low density, these may be additional factors that contributes to underestimated ΔF in the imputed sets. To find the effect of imputation when measuring inbreeding, there is a need to test new datasets imputed up to a high density with high density and no removal of low MAF SNP to be able to detect the actual effect of imputation. Also, it would be preferable to use imputation software that utilizes a pedigree in addition to genomic data.

An assumption which was made here to estimate Ne was that homozygosity was increasing over time due to the inbreeding, and thus that heterozygosity was decreasing. The latter requires that the heterozygosity was much higher in the past, and has been decreasing since. This assumptions seemed justified for F_{Hom}, since SNP were generally old mutations, and historical effective population sizes were very large in cattle [5]. For F_{ROH}, Hayes et al. [2] showed that the current chromosome segment homozygosity reflected effective population sizes 1/(2c) generations ago, where c was the size of the segment in Morgans. ROH was detected with minimum length of 0.5

and 2 Mb, which yielded c values of .005 and 0.02, respectively (assuming an approximate genetic distance 0.01 Morgans/Mb). Thus, our ROH's came from common ancestors 100 and 25 generations ago. The past reductions in Ne may be not so large during the last 25 generations, which may cause a reduced loss of heterozygosity (the population became closer to a steady state, where F_{ROH} was constant), explaining the larger Ne estimates when F_{ROH} was used, especially with segments > 2 Mb. On the other hand, a major population admixture event occurred in the Norwegian Red population in the '60 and '70. This means that old bulls may have shown relatively high degrees of heterozygosity due to these crossing events, whereas in the current bulls the original lines may meet again in an individual causing relatively high degrees of homozygosity. That way, the loss of heterozygosity may have been inflated over the studied period due to an early population admixture event.

In summary, it is recommended to estimate individual inbreeding by utilizing observed homozygosity, which accounts better for the increase in homozygosity than ROH. As for ROH, the individual value of observed homozygosity will become more precise as SNP density increases, but for calculation of ΔF a density of 54K suffices. When regressing on CGE, the effective population size was only 57.5 animals; 1/3 of that obtained traditionally when regressing on year of birth. These results were obtained only with bulls, but should also be relevant for the entire population, following Woolliams, Mantysaari [27]. Further, the main results were obtained in a restricted sample of the population of bulls, and should be recalculated as additional high-density data becomes available.

Conclusions

It was not only possible to measure Ne and ΔF by using either observed homozygosity or ROH, but it also seemed to result in more accurate estimates than pedigree because the pedigree data suffered of a threshold effect. Preference was given to observed homozygosity over ROH because it produced stable results of ΔF, even at a density of 53K. ROH gained more from an increasing SNP density, and produced results intermediate to those from observed homozygosity and pedigree. In this population, rate of inbreeding should be estimated from regressing ln(1-F_{Hom}) on CGE, rather than by year of birth. Further, low MAF SNP should not be removed from the data. Imputation programs that do not utilize pedigree, may cause additional error detecting homozygosities and should be investigated further.

Competing interests

The authors declare that they have no competing interests.

Author’s contributions

All authors designed the study, interpreted the findings and revised the manuscript. BH and JAW ran the calculations. BH, JAW, TM and GK analyzed the results. BH drafted the manuscript. JAW, TM, DIV and GK co-wrote the manuscript.

Acknowledgments

We would like to thank the Norwegian University of Life Sciences for funding this project. Thank you, Geno, for sharing pedigree files and genotyping data, especially Morten Svendsen and Trygve Roger Solberg for being very helpful finding the information needed. Also, a big thanks to Solomon Antwi Boison and Harald Grove for contributing on genotyping quality controls, Sreten Andonov for introducing Pedig and Trygve Flathen for solving programming issues.

References

1. Wright S. Coefficients of Inbreeding and Relationship. The American Naturalist. 1922;56(645):330-8. doi:10.2307/2456273.

2. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research. 2003;13(4):635-43. doi:10.1101/gr.387103.

3. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME et al. Recent human effective population size estimated from linkage disequilibrium. Genome Research. 2007;17(4):520-6. doi:10.1101/gr.6023607.

4. Corbin LJ, Liu AYH, Bishop SC, Woolliams JA. Estimation of historical effective population size using linkage disequilibria with marker data. Journal of Animal Breeding and Genetics. 2012;129(4):257-70. doi:10.1111/j.1439-0388.2012.01003.x.

5. MacLeod IM, Meuwissen THE, Hayes BJ, Goddard ME. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genetics Research. 2009;91(6):413-26. doi:10.1017/s0016672309990358.

6. Saura M, Fernandez A, Rodriguez MC, Toro MA, Barragan C, Fernandez AI et al. Genome-Wide Estimates of Coancestry and Inbreeding in a Closed Herd of Ancient Iberian Pigs. Plos One. 2013;8(10). doi:10.1371/journal.pone.0078314.

7. Bjelland DW, Weigel KA, Vukasinovic N, Nkrumah JD. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. Journal of Dairy Science. 2013;96(7):4697-706. doi:10.3168/jds.2012-6435.

8. Silió L, Rodríguez MC, Fernández A, Barragán C, Benítez R, Óvilo C et al. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. Journal of Animal Breeding and Genetics. 2013:n/a-n/a. doi:10.1111/jbg.12031.

9. Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. Am J Hum Genet. 1999;65(6):1493-500. doi:10.1086/302661.

10. McQuillan R, Leutenegger A, Abdel-Rahman R, Franklin C, Pericic M, Barac-Lauc L et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83(3):359 - 72.

11. Hillestad B, Woolliams JA, Boison SA, Grove H, Meuwissen T, Våge DI et al. Detection of runs of homozygosity in Norwegian Red: Density, critera and genotyping quality control. Submitted to GSE. 2015.

12. Illumina. BovineSNP50 Genotyping BeadChip. 2012. http://res.illumina.com/documents/products/datasheets/datasheet_bovine_snp5o.pdf.

13. Affymetrix. Data Sheet: Affymetrix Targeted Genotyping Bovine 25K SNP Service. Affrymetrix, Inc. 2007. http://www.affymetrix.com/support/technical/byproduct.affx?product=bo-25ksnp.

14. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084-97. doi:10.1086/521987.

15. Illumina. Bovine HD Genotyping BeadChip Datasheet. 2012. http://res.illumina.com/documents/products/datasheets/datasheet_bovinehd.pdf.

16. Boichard D. Pedig: a fortran package for pedigree analysis suited to larger populations. 7th World Congress on Genetics Applied to Livestock Production; 19th-23rd of August; Montepellier2002. p. paper 28-13.

17. Maccluer JW, Boyce AJ, Dyke B, Weitkamp LR, Pfennig DW, Parsons CJ. Inbreeding and pedigree structure in standardbred horses. Journal of Heredity. 1983;74(6):394-9.

18. Stranden I. RelaX2. Version 1.10 Update 10/07 ed. FIN-31600 Jokioinen, Finland: MTT, Biometrical Genetics, Agrifood Research Finland; 2006.

19. Meuwissen THE, Luo Z. Computing inbreeding coefficients in large populations. Genet Sel Evol. 1992;24(4):305-13. doi:10.1051/gse:19920402.

20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 812007.

21. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4 ed. England: Pearson Education Limited; 1996.

22. Edriss V, Guldbrandtsen B, Lund MS, Su G. Effect of marker-data editing on the accuracy of genomic prediction. Journal of Animal Breeding and Genetics. 2013;130(2):128-35. doi:10.1111/j.1439-0388.2012.01015.x.

23. Rodríguez-Ramilo ST, Fernández J, Toro MA, Hernández D, Villanueva B, editors. Genome-wide estimates of effective population size in the Spanish Holstein population. WCGALP; 2014; Vancouver, Canada.

24. Gómez-Romano F, Sölkner J, Villanueva B, Mézáros G, Cara MARd, O'Brien AMP et al., editors. Genomic estimates of inbreeding and coancestry in Austrian Brown Swiss cattle. WCGALP; 2014; Vancouver, Canada.

25. Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME. Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. Genetics. 2011;189(1):317-U1028. doi:10.1534/genetics.111.128082.

26. Schrooten C, Dassonneville R, Ducrocq V, Brondum RF, Lund MS, Chen J et al. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genet Sel Evol. 2014;46. doi:10.1186/1297-9686-46-10.

27. Woolliams JA, Mantysaari EA. Genetic contributions of Finnish Ayrshire bulls over 4 generations. Animal Science. 1995;61:177-87.

Table 1: Datasets used to measure inbreeding

Subsets varying in SNP density and genotyping quality control (HDQ and IQ, with additional pruning as described in Material and Methods) used to find rate of inbreeding ΔF and effective population size (Ne) in Norwegian Red.

Density	Exact # of SNP	SNP pr Kb	# of animals
	HDQ		
53K	53,129	0.0177	375
94K	94,452	0.0315	375
224K	223,890	0.0746	375
531K	530,706	0.1769	375
708K	707,609	0.2359	375
	IQ		
$48K_{GS}$	48,249	0.0161	2,372
$540K_{GS}$	539,665	0.1799	375

Table 2: PLINK constraints to detect ROH

Criteria used for identifying ROH in PLINK using 5 Mb sliding windows for different SNP densities. The SNP densities arose from 2 different quality control methods (HDQ and IQ) as described in Materials and Methods. For all ROH identified no heterozygote SNP was allowed and the minimum length was required to be > 500 Kb, except when minimum length was tested at > 2 Mb.

SNP density	SNP/5Mb	PLINK constraints			
		Max. # missing SNP/window	Per ROH		
			Min # SNP	Max gap (Kb)	Max avg. Kb/ SNP
		HDQ			
53K	88.5	1	9	1,000	150
94K	157.4	1	16	1,000	150
224K	373.2	2	37	250	50
531K	884.5	3	88	250	50
708K	1,179.3	3	118	250	50
		IQ			
$48K_{GS}$	80.4	1	8	1,000	150
$540K_{GS}$	899.4	3	90	250	50

Table 3: Rate of inbreeding and effective population size based on ROH, observed homozygosity and pedigree using different SNP densities

Rate of inbreeding (ΔF) and effective population size (Ne) estimated on 375 Norwegian Red bulls born between 1975 and 2004, regressed by year of birth or complete generation equivalent (CGE). The estimates are estimated from pedigree, runs of homozygosity (ROH) and observed homozygosity, when genomic data ranged between 53-708K SNP densities from HDQ quality controls as described in Material and Methods. ROH criteria are described in Table 2. ΔF and standard errors are scaled by 10^3.

Approach		By year			By CGE		
		ΔF (se)	R^2	Ne	ΔF (se)	R^2	Ne
Pedigree	F_{Ped}	2.57 (0.52)	0.06	194.6	4.17 (0.56)	0.13	119.9
			HDQ				
ROH	53K	3.23 (0.98)	0.03	154.8	6.19 (1.07)	0.08	80.8
	94K	3.46 (1.00)	0.03	144.5	6.66 (1.09)	0.09	75.1
	224K	3.85 (1.00)	0.04	129.9	7.12 (1.09)	0.10	70.2
	531K	3.75 (1.01)	0.04	133.5	7.06 (1.09)	0.10	70.8
	708K	3.69 (1.00)	0.03	135.7	6.96 (1.09)	0.10	71.8
Observed homozygosity	53K	5.37 (1.11)	0.06	93.2	8.60 (1.21)	0.12	58.1
	94K	5.33 (1.10)	0.06	93.9	8.65 (1.20)	0.12	57.8
	224K	5.40 (1.10)	0.06	92.6	8.62 (1.20)	0.12	58.0

531K	5.45 (1.11)	0.06	91.8	8.71 (1.20)	0.12	57.4
708K	5.40 (1.10)	0.06	92.6	8.69 (1.20)	0.12	57.5

Table 4: Basic statistics for inbreeding coefficients using different SNP densities

Average values and correlations of F-values from pedigree (Ped), runs of homozygosity (ROH) and observed homozygosity (Hom) using different SNP densities between 53K and 708K, raising the minimum length of ROH from 0.5 to 2 Mb and varying in genotyping quality controls (HDQ and IQ) as described in Material and Methods. ROH criteria are described in Table 2. The exact same animals were included in all datasets, a total of 375 Norwegian Red bulls born between 1975 and 2004, with an average F_{Ped} of 0.020 and a complete generation equivalent (CGE) of 7.48.

Density	F_{ROH}	F_{Hom}	Cor(F_{Hom},F_{ROH})	Cor(F_{Ped},F_{ROH})	Cor(F_{Ped},F_{Hom})
			HDQ		
53K	0.062	0.646	0.876	0.542	0.508
94K	0.071	0.645	0.892	0.540	0.516
224K	0.095	0.646	0.913	0.538	0.510
531K	0.095	0.646	0.913	0.535	0.511
708K	0.092	0.646	0.913	0.534	0.512
			Minimum length > 2 Mb		
53K	0.062	0.646	0.876	0.542	0.508
708K	0.059	0.645	0.895	0.539	0.512
			IQ		
$48K_{GS}$	0.041	0.629	0.902	0.569	0.487
$540K_{GS}$	0.037	0.610	0.921	0.544	0.534

Table 5: Relationship between genomic and pedigree based inbreeding coefficients using different SNP densities

Slopes and standard errors of the regression *ln*(1-F_y)=μ+β**ln*(1-F_{Ped}), where F_y is either F_{ROH} of F_{Hom}, μ=*ln*(1-F_{Pop}) and F_{Pop} is a population mode of F. F is the individual inbreeding coefficient, Ped is pedigree, ROH is runs of homozygosity and Hom equals observed homozygosity. The expected relationship of F_y- and F_{Ped}-values was exploited using different SNP-densities between 53K and 708K, raising the minimum length of ROH from 0.5 to 2 Mb and varying in genotyping quality controls (HDQ and IQ) as described in Material and Methods. ROH criteria are described in Table 2. This was done on the exact same animals in all datasets, a total of 375 Norwegian Red bulls born between 1975 and 2004.

Density	F_{ROH}		F_{Hom}	
	β	se	β	se
SNP densities with HDQ				
53K	1.01	0.08	1.09	0.10
94K	1.02	0.08	1.10	0.10
224K	1.03	0.08	1.09	0.10
531K	1.02	0.08	1.09	0.10
708K	1.02	0.08	1.09	0.10
Minimum length > 2 Mb with HDQ				
53K	1.01	0.08	1.09	0.10
708K	0.96	0.08	1.09	0.10
IQ				
$48K_{GS}$	0.83	0.06	0.90	0.08
$540K_{GS}$	0.48	0.04	1.00	0.08

Table 6: Rate of inbreeding and effective population size based on ROH, observed homozygosity and pedigree using different constraints

Rate of inbreeding (ΔF) and effective population size (Ne) estimated on 375 Norwegian Red born between 1975 and 2004, regressed by year of birth or CGE. The estimates are made on pedigree, runs of homozygosity (ROH) and observed homozygosity, by altering the minimum length of ROH between 0.5 and 2 Mb and by varying genotyping quality controls (HDQ and IQ) as described in Material and Methods. ROH criteria are described in Table 2. ΔF and standard errors are scaled by 10^3.

Approach		By year		By CGE	
		ΔF (se)	Ne	ΔF (se)	Ne
Pedigree	F_{Ped}	2.57 (0.52)	194.6	4.17 (0.56)	119.9
	Minimum length > 2 Mb with HDQ				
ROH	53K	3.24 (0.98)	154.2	6.07 (1.05)	82.4
	708K	3.22 (0.94)	155.5	5.95 (1.01)	84.0
	IQ				
ROH	$48K_{GS}$	2.22 (0.77)	225.2	4.20 (0.83)	119.2
	$540K_{GS}$	1.74 (0.46)	297.9	3.07 (0.49)	162.8
Observed homozygosity	$48K_{GS}$	2.99 (0.98)	167.4	5.02 (1.05)	99.7
	$540K_{GS}$	4.30 (0.98)	116.2	6.84 (1.05)	73.1

Table 7: Basic statistics for inbreeding coefficients using imputed genotypes

Average values and correlations of F-values from pedigree, runs of homozygosity (ROH) and observed homozygosity (Hom) in imputed and non-imputed datasets for Norwegian Red bulls born between 1975 and 2009. Average F_{Ped} equal to 0.022 and complete generation interval (CGE) of 8.71. All sets ends up with a density of 48K after genotyping quality controls and imputation, adding missing SNP from either the 25K or the 54K chip. ROH criteria are described in Table 2.

Original genotyping	# of animals	F_{ROH}	F_{Hom}	Cor(F_{Hom},F_{ROH})	Cor(F_{Ped},F_{ROH})	Cor(F_{Ped},F_{Hom})
Both (25K and 54K)	448	0.040	0.628	0.888	0.568	0.493
25K	1,704	0.039	0.630	0.888	0.568	0.490
54K	1,116	0.044	0.631	0.795	0.615	0.398

Table 8: Relationship between genomic and pedigree based inbreeding coefficients using imputed genotypes

Slopes and standard errors of the regression $ln(1-F_y)=\mu+\beta*ln(1-F_{Ped})$, where F_y is either F_{ROH} or F_{HOM}, $\mu=ln(1-F_{Pop})$ and F_{Pop} is a population mode of F, where F is the individual inbreeding coefficient, Ped is pedigree, ROH is runs of homozygosity and Hom equals observed homozygosity. The expected relationship of F_y and F_{Ped} was exploited using imputed and non-imputed subsets. All sets ends up with a density of 48K after IQ genotyping quality controls as described in Material and Methods and imputation with missing SNP from either the 25K or the 54K chip. ROH criteria are described in Table 2.

Original	F_{ROH}		F_{Hom}	
genotyping	β	se	β	se
Both (25 and 54K)	0.83	0.06	0.92	0.08
25K	0.79	0.03	0.89	0.04
54K	0.85	0.03	0.83	0.05

Table 9: Rate of inbreeding and effective population size based on ROH, observed homozygosity and pedigree using imputed genotypes

Rate of inbreeding (ΔF) and effective population size (Ne) estimated on Norwegian Red bulls born between 1975 and 2009 in imputed and non-imputed datasets. The estimates were utilized on inbreeding coefficients from pedigree (Ped), runs of homozygosity (ROH) and observed homozygosity (Hom), respectively, regressed by year of birth or by complete generation equivalent (CGE). All subsets ends up with a density of 48K after IQ genotyping quality controls (as described Material and Methods) and imputation with missing SNP from either the 25K or the 54K chip. ROH criteria are described in Table 2. ΔF and standard errors are scaled by 10^3.

Original genotyping	F_{Ped}		F_{ROH}		F_{Hom}	
	ΔF (se)	Ne	ΔF (se)	Ne	ΔF (se)	Ne
By year						
Both (25 and 54K)	2.51 (0.50)	199.2	1.66 (0.73)	301.9	1.85 (0.99)	270.3
25K	2.42 (0.26)	206.7	1.12 (0.38)	448.4	0.97 (0.51)	516.5
54K	5.00 (0.30)	100.0	3.87 (0.44)	129.1	2.89 (0.60)	172.9
By CGE						
Both (25 and 54K)	3.79 (0.55)	131.8	3.16 (0.82)	158.3	3.87 (1.11)	129.2
25K	3.39 (0.28)	147.4	1.96 (0.41)	255.4	2.20 (0.56)	227.1

54K	5.47 (0.29)	91.5	4.67 (0.42)	107.0	3.67 (0.57)	136.1

Figure 1: Matrix plot of year of birth, complete generation equivalent, ln(1-F_{Ped}), ln(1-F_{ROH}) and ln(1-F_{Hom})

Regression matrix, with ordinary (Regress) and locally weighted least-squares (Lowess) regression as well as data points, of year of birth, complete generation equivalent (CGE) and ln(1-F_{Ped}), ln(1-F_{ROH}) and ln(1-F_{Hom}) in 375 Norwegian Red bulls genotyped with a 708K Illumina HD-panel. The genotypes had HDQ quality controls as described in Material and Methods. ROH criteria are described in Table 2.

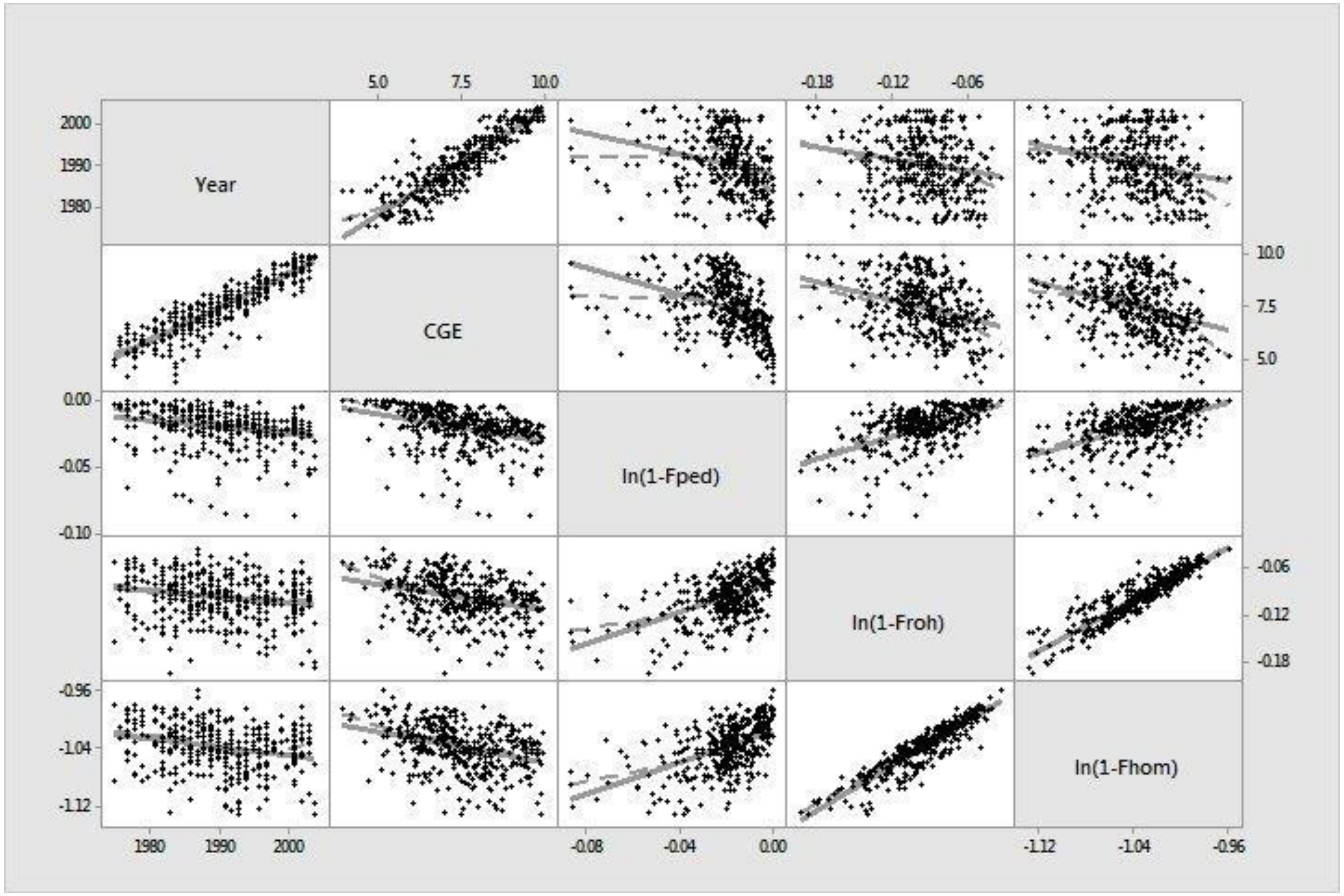
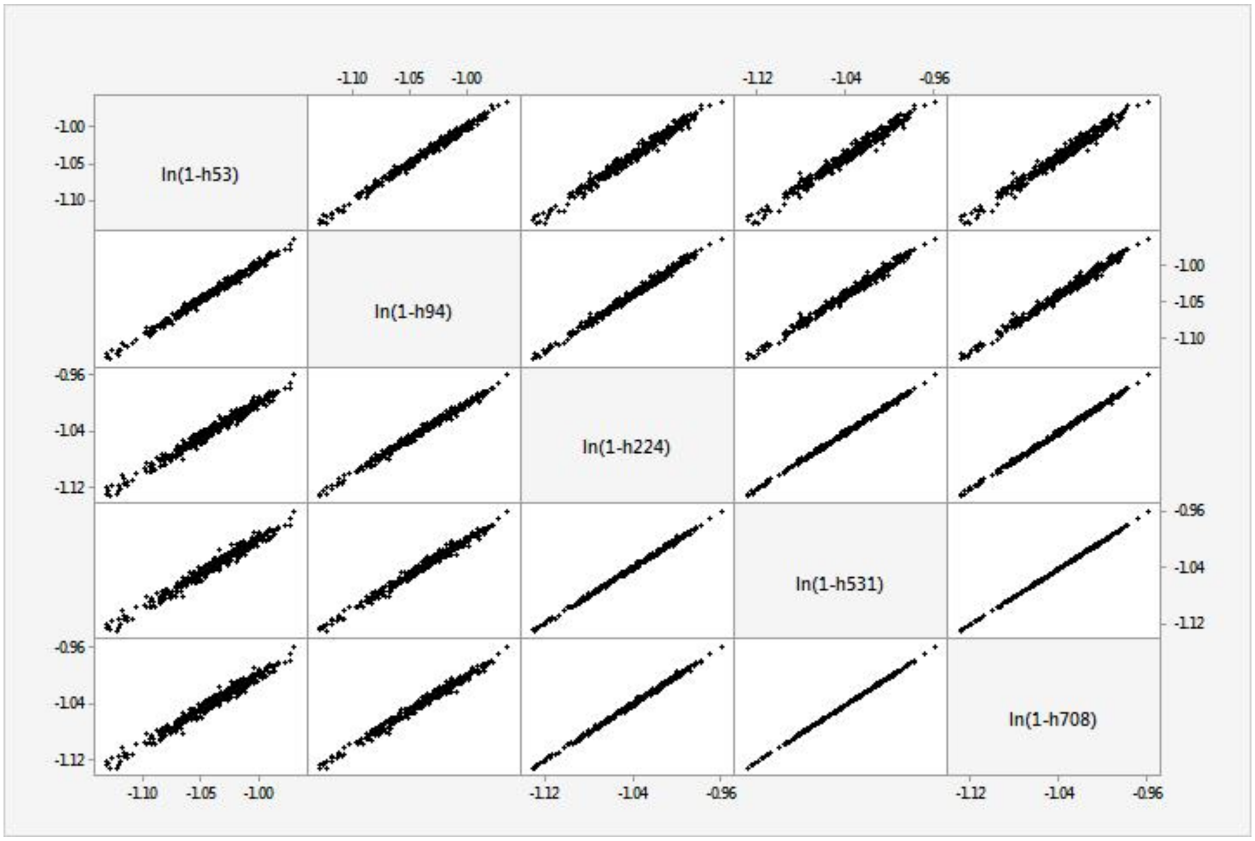

Year
CGE
ln(1-Fped)
ln(1-Froh)
ln(1-Fhom)
5.0
7.5
10.0
-0.18
-0.12
-0.06
2000
1990
1980
10.0
7.5
5.0
0.00
-0.05
-0.10
-0.06
-0.12
-0.18
-0.96
-1.04
-1.12
1980
1990
2000
-0.08
-0.04
0.00
-1.12
-1.04
-0.96

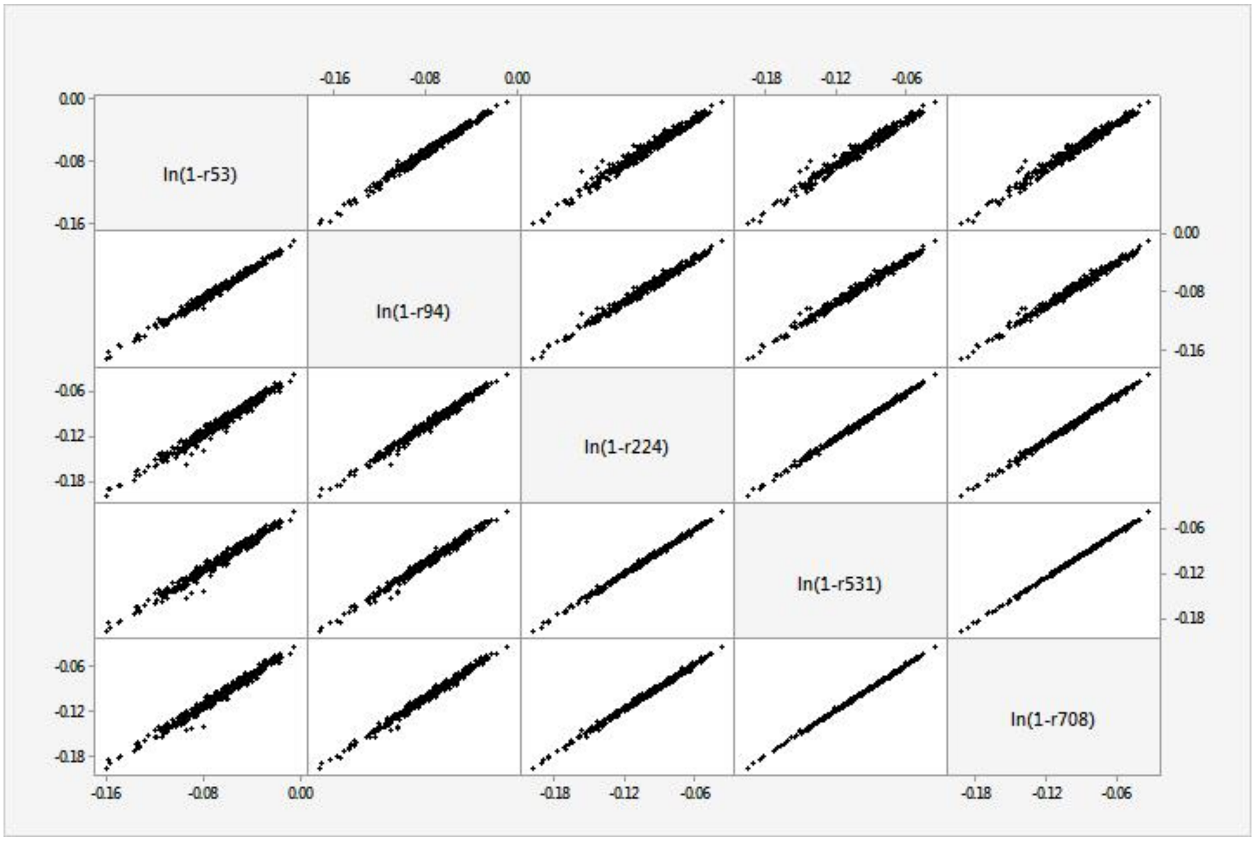
Figure 1

Matrix plot of ln(1-F_{Hom}) using different SNP densities



S1: Matrix plot of F_{Hom} utilized from different SNP densities in 375 Norwegian Red bulls genotyped with a 708K Illumina HD-panel. The genotypes had HDQ quality controls as described in Material and Methods. The plot illustrated how an increased SNP density removed error.

Matrix plot of ln(1-F_{ROH}) using different SNP densities



S2: Matrix plot of F_{ROH} utilized from different SNP densities in 375 Norwegian Red bulls genotyped with a 708K Illumina HD-panel. The genotypes had HDQ quality controls as described in Material and Methods. The plot illustrated how an increased SNP density removed error. ROH criteria are described in Table 2.

Paper III

Screening for selection signatures in Norwegian Red

Borghild Hillestad, John Arthur Woolliams, Solomon Antwi Boison, Dag Inge Våge, Gunnar Klemetsdal

Screening for selection signatures in Norwegian Red

Borghild Hillestad[1], John Arthur Woolliams[1,2], Solomon Antwi Boison[3], Dag Inge Våge[1,4], Gunnar Klemetsdal[1]

[1]Department of Animal and Aquacultural Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, N-1432 Ås, Norway

[2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

[3]University of Natural Resources and Life Sciences Vienna, Department of Sustainable Agricultural Systems, Division of Livestock Sciences, Gregor Mendel Str. 33, A-1180 Vienna, Austria

[4]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, N-1432 Ås, Norway

Borghild Hillestad borghildhillestad@gmail.com

John Arthur Woolliams john.woolliams@roslin.ed.ac.uk

Solomon Antwi Boison soloboan@yahoo.com

Dag Inge Våge daginge.vage@nmbu.no

Gunnar Klemetsdal gunnar.klemetsdal@nmbu.no

Corresponding author: Borghild Hillestad

Abstract

Background: Due to the possibility of estimating individual inbreeding using genomic data, narrowing down the rate of inbreeding on a segmental level is of interest to map where on the genome inbreeding occurs. The object of this study was to locate segments exposed to inbreeding, map the rate of inbreeding on a segmental level and find selection signatures using ROH in Norwegian Red.

Material and Methods: The dataset contained 384 Norwegian Red bulls genotyped with the Illumina HD-panel containing 777K SNP-markers. After genotyping controls, 381 animals born between 1971 and 2004 and 708,609 SNP remained to estimate individual inbreeding coefficients (F-values) based on observed homozygosity on a chromosomal level and by runs of homozygosity (ROH) on a positional levels.

Results: By regressing the individual F-values on complete generation equivalent (CGE), some chromosomes were found to be more inbred than others. The bovine chromosomes 5, 14 and 24 were estimated to have the lowest Ne, ranging between 22.6 and 34.2. Positional F-values on each SNP were made from ROH, with the highest values on BTA 1, 5, 7, 14 and 22. With logistic regression of ROH status on CGE and ROH-plots, ongoing selective sweeps were identified on BTA 5, 6, 12 and 24. Footprints like historical sweeps and deserts of missing SNP were also observed.

Conclusions: ROH is an effective screening method for selection signatures in the absence of phenotypes, and allowed to discriminate between ongoing and historical selective sweeps.

Keywords: Runs of homozygosity (ROH), genomic inbreeding, observed homozygosity, selection signatures, cattle

Background

Inbreeding is associated with inbreeding depression, and the depression is synonymous with increased risk of homozygous recessives [1]. The individual inbreeding coefficient (**F**) represent the strength of inbreeding and is defined as the probability that two alleles in an individual locus are identical by descent (**IBD**). For a long time the F-values have been estimated using pedigree information in livestock production, but lately several studies have calculated inbreeding by including genomic data [2-5]. The combination of both pedigree and genomic data seemed to provide better estimates of inbreeding than by pedigree or genomic data separately. Hillestad et al. [6] found observed homozygosity and runs of homozygosity (**ROH**) to be suitable methods measuring rate of inbreeding (**ΔF**), by regressing ln(1-F) on the complete generation equivalent (**CGE**) (i.e. the number of generations an individual could be traced back with complete pedigree information).

The availability of genomic data also makes it possible to locate where inbreeding is manifested at the genome. By mapping homozygosity over time, selection signatures like historical and ongoing selective sweeps may be detected. Selective sweep is an event that reduce the genetic variation of a region, due to the positive selection for a new favorable variant that sweeps all other variants away [7]. Thus, by observing change of segmental homozygosity over time, selective sweeps could be detected. A high rate of change in positional homozygosity could indicate segments under strong selection [8]. ROH has the advantage of detecting segmental homozygosity. Each inherited segment would be split into shorter segments from one generation to the next, hence reduce the length of the original segments. The rate of change over time based on ROH, as a function of position can therefore be used to detect selection signatures without any use of phenotypic information.

Even though the mating of two animals will result in inbred offspring if their parents are related, they may not necessarily be inbred at the same areas on the genome. By knowing how inbreeding is distributed in each animal genome, breeding could be further optimized. The object of this study is therefore to map the rate of inbreeding on a chromosomal and segmental level using observed homozygosity and ROH, and identify selection signatures in Norwegian Red.

Materials and Methods

Genotypes

In this study, 384 Norwegian Red bulls born between 1971 and 2004 were genotyped with the Illumina HD-panel, containing 777,962 SNP-markers, covering 2.51 Gb of the 3 Gb large genome. After genotyping, the marker data passed through several stages of quality controls to exclude markers on sex-linked chromosomes, call rate per SNP > 90 % (individual SNP score missing if GenCall score < 0.7) and deviation from Hardy-Weinberg ($P > 10^{-6}$). Three animals were removed for having genotypes for fewer than 95 % of loci. This resulted in the retention of 707,609 SNP and 381 animals.

Chromosome wise inbreeding estimates

To identify the most inbred chromosomes, ΔF and Ne at each chromosome were estimated. First, for each individual on each chromosome, an individual inbreeding coefficient (**F_{Hom*j*}**) was estimated based on the amount of observed homozygous SNP on that chromosome:

$$F_{\mathrm{Hom}_j} = \mathrm{O(Hom)}_j \,/\, \mathrm{N(NM)}_j \qquad (1)$$

where $\mathrm{N(NM)_j}$ was defined as the number of non-missing genotypes at chromosome *j* and $\mathrm{O(Hom)}_{\mathrm{j}}$ the amount of observed homozygosity at the corresponding chromosome.

To estimate the chromosomal rate of inbreeding, individual values of $\ln(1\text{-}F_{\mathrm{Hom}j})$ were regressed on the complete generation equivalent (**CGE**). CGE was estimated from pedigree that extended as far back as 1875, using Pedig [9] based on the equation of Maccluer et al. [10]:

$$\mathrm{CGE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{j=1}}^{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{n_j}}\frac{1}{2^{\mathrm{g_{ij}}}} \qquad (2)$$

Here *N* refers to number of genotyped animals; $n_{j,}$ the total number of ancestor of animal *j* in the population in this study; and g_{ij}, the number of generations between *j* and its ancestor *i*. The CGE were traced back no more than 20 generations per individual due to limitations in Pedig.

Formally, the regression equation used to estimate ΔF followed the derivation of Hillestad et al. [6]:

$$\begin{aligned} y_i &= \mu + \beta t_i + \mathrm{e}_i \\ \Delta F &= 1 - \mathrm{e}^{\beta} \end{aligned} \qquad (3)$$

where y_i referred to ln(1-$F_{\mathrm{Hom}j}$) of individual *i* and t_i to the CGE of individual *i*. The slope was utilized to calculated ΔF, and finally chromosomal Ne was obtained by the following equation:

$$\mathrm{Ne} = \frac{1}{2\Delta \mathrm{F}} \qquad (4)$$

As in Hillestad et al. [6], six bulls were deleted from the dataset; those born before 1975 and one bull with high leverage when regressing across chromosomal genomic heterozygosity on pedigree heterozygosity, leaving 375 bulls for analysis.

Utilizing ROH data

ROH were identified with PLINK 1.07 [11]. PLINK operates with sliding windows of 5,000 Kb, determining homozygosity at each window. When using a 708K dataset, there is an average of 1,179.3 SNP present in each window. Based on Hillestad et al. [12], the following criteria were set to define a ROH: (i) The minimum number of adjacent homozygous SNP loci were set to 118, based on the fact that on average 118 SNP would be present on a 500 Kb ROH at a 708K density on a 3 Gb genome; (ii) no heterozygous SNP were allowed within a ROH; (iii) three missing SNP were allowed per window; (iv) maximum physical distance between adjacent SNP within a ROH (maximum gap length) were set to 250 Kb and (v) the minimum average density of SNP within a ROH was set to 50 Kb.

A positional inbreeding coefficient (**F$_{\mathbf{j}}$**) for each SNP *j* were estimated by the following formula:

$$F_j = \frac{\sum_{i=1}^{N} s_{ij}}{N} \qquad (5)$$

where s_{ij} was the status of the locus, whether it is within a ROH or not (1 or 0) for animal *i*, and *N* is the total number of animals with genomic data. Two different F$_j$ were estimated for each SNP: (i) One with a minimum length for ROH of 0.5 Mb (**F$_{\mathbf{j(0.5)}}$**); (ii) and a second with minimum length for ROH of 2 Mb (**F$_{\mathbf{j(2)}}$**).

Further, the rate of change of s$_{ij}$ per generation (CGE) was estimated for each SNP by logistic regression and by use of the following likelihood function:

$$\begin{aligned} &L(\beta_j) = \prod_{i=1}^{N} Bernoulli(p_{ij}) \\ &p_{ij} = \frac{\exp(\eta_{ij})}{1-\exp(\eta_{ij})} \\ &\eta_{ij} = [\eta_{1j}.....\eta_{Nj}]' \\ &\log it(p_{ij}) = \eta_{ij} = \mu_j + \beta_j t_i \end{aligned} \qquad (6)$$

where μ was the intercept and β the slope on position *j*, and *t* the CGE in individual *i*, respectively.

The slope of change of s_{ij} was plotted chromosome wise, and segments with a –log(p) > 4 were defined as significant. Further, visualization of the change of ROH over time was obtained by plotting all detected ROH in each animal chromosome wise, ordered by date of birth.

Results

Chromosomal inbreeding

When chromosome wise ΔF and Ne were estimated from observed homozygosity regressed on CGE on each chromosome, the regressions were found nominal significant ($p < 0.05$) at BTA 5, 6, 9, 11, 14, 15, 16, 20, 21, 23 and 24 (Table1). BTA 5, 14 and 24 were also found Bonferroni significant. Chromosome wise, the estimates of Ne ranges from 22.6 on BTA 24 to 418 on BTA 22, as compared to the average autosomal estimate of 57.5 [6].

ROH estimates

Positional F from ROH. By raising minimum lengths of ROH to 2 Mb, fewer ROH were detected than with a 0.5 Mb threshold (Table 2). The longest ROH detected reached over 58 Mb. Per animal, the lowest number of segments detected was 1 ROH for a minimum length of ROH of 2 Mb, in contrast to 72 ROH of 0.5 Mb threshold. This questioned the credibility of the estimated inbreeding measurements when such a high threshold was set for minimum length detecting ROH.

Positional F for a minimum length of 0.5 Mb ($F_{j(0.5)}$) versus 2 Mb ($F_{j(2)}$) are shown in Figure 1. The highest values of $F_{j(0.5)}$ were found on chromosomes 1, 5, 7, 14, and 22, indicating much homozygosity on these chromosomes. The homozygosity level did not correspond with the chromosomal rate of inbreeding being most expressed on BTA 5 and 14 and only minor on BTA 1 and 22 (Table 1).

SNP wise rate of ROH over time. For the rate of change of status (β_j), a total of 4 segments on BTA 5, 6, 12 and 24 were found significant by having a $-\log(p) > 4$ (Figure 2). At the peaked value of the test statistics, β_j was also in general somewhat enlarged. The identified segments were: (i) A segment on 70-95 Mb in BTA 5, (ii) 45-64 Mb on BTA 6, (iii) 10-20 Mb on BTA 12, and (iv) 10-20 Mb on BTA, for which some detailed ROH information is given in Table 3. In general, the identified segments had some extremely long ROH, and the longest ROH of the entire genome on this dataset began at the second half of the segment on BTA 6 reaching over 58.7 Mb, which appeared in two different animals with approximately the same start and stop location, indicating similar haplotype.

The distribution of ROH in each animal was also plotted ordered by year of birth and ID number, where the oldest animals were placed closest to the horizontal line and the youngest to the top of the plot, illustrating the dynamics of ROH changed over time, from 1971 to 2004 (Figure 3). It was also confirmed that the frequency of ROH were increasing over time at the peaked –log(p) values of Figure 2 on BTA 5, 6, 12 and 24, indicating ongoing selective sweeps.

The position of the well-known DGAT1 at 1.8 Mb in BTA 14 [13, 14] did show an excess of ROH, but did neither show any sweep nor a total fixation. However, this chromosome did have a long fixed haplotype from 24-25 Mb, illustrating a historical sweep. In BTA 6 at 52-53 Mb, Figure 2 showed a drop of –log(p) from approximately 3 to 0, saying that no change of ROH frequency was occurring at the area over time. Also, Figure 3 showed a high frequency of ROH at this area, indicating a historical selective sweep. At the same time an ongoing selective sweep have been indicated between 45-65 Mb, implying that the area had a mixture of two events: both an ongoing and a historical sweep.

An additional event that became visible through ROH-plots were deserts of missing SNP markers, for instance at BTA 12 around 75 Mb. This gap was so big that ROH were not allowed to be detected there or nearby.

Discussion

In this paper we mapped inbreeding on a chromosomal and segmental level, and several chromosomes stood out with a significantly lower Ne compared to others. This implies that some chromosomes were more inbred than others. ROH seemed to be a good screening method to identify selection signatures without any phenotypes available. It was demonstrated that further inference could be obtained by plotting individual ROH over time on a segmental level, which allowed to discriminate between historical from ongoing selective sweeps.

When individuals were plotted on a time scale in ROH-plots, ongoing selective sweeps were visualized, confirming the peaked plotting of the test statistics from logistic regression. Further, ROH plotting made it possible to make inference to historical sweeps, because low MAF SNP were not removed when detecting ROH. Thus, the increased homozygosity around a core haplotype would be visible as long as the homozygous segment was larger than the minimum length defined for ROH. Many methods have been developed to detect selection signatures, among other methods based on linkage disequilibrium (**LD**) [15]. One challenge with LD-based tests are the dependency on allele frequencies to the core haplotype. When an allele reaches fixation at this core, the frequency approach zero and the method reduces its power to detect selection signature. This did not happen when ROH-plots were used, but was a weakness of the logistic regression approach that heavily relied on the access of genomic data over a long period of time.

Due to the long generation interval in cattle, a study including more animals and larger time span would be preferable to obtain a more detailed picture of chromosomal changes due to selection. Selection signatures are an evolutionary process, and a selective sweep may not be visible if only a short period of time is studied [16]. A so called hard sweep is created when a new favorable allele sweeps off the genetic variation of the loci, while the allele causing a soft sweep has been among the genetic variation for a longer time, but recently become advantageous. Thus, a hard sweep would be easier to detect, and sweep off genetic variation sooner than a soft sweep that will sweep more gradually. With a generation interval of 5 years gaining only 4.6 generations within this dataset, this process will span over a long period in years, and if the segment of interest is not yet defined an even broader perspective is needed.

Regarding historical sweeps, BTA 14 stood out with high levels of F_j and a low chromosomal Ne based on observed homozygosity, but did not stand out with high β_{ij} or -log(p) values, terminating the possibility for any ongoing selective sweeps at the chromosome. Hillestad et al. [12] reported that BTA14 contained 23.9 % SNP with MAF < 0.01 on the Illumina HD-panel. Since this chromosome contained most low MAF SNP next after BTA8 in this population, this supports the signals of a chromosome containing many fixed haplotypes. Thus, by keeping low MAF SNP both ongoing and historical selective sweep are detected. Fixed haplotypes are a natural consequence of selection, because one haplotype variant are selected for. BTA 14 contains gene variants influencing many economical important traits for both milk and beef cattle breeds, and has been a chromosome under study and selection for a long time [17]. One of the genes at BTA 14 is the well-known DGAT1 affecting milk fatty acid [14]. Even though ROH was detected in some animals at this position, there were no clear signals of strong selection at this area, and the gene may not be segregating in Norwegian Red, an assumption also supported by Karlengen et al. [18]. On the other hand, a QTL of protein yield was reported in Holstein by Ashwell et al. [19] at BTA 14 at 24.7-27.3 Mb, and could be the reason of the historical sweep at 25 Mb on BTA 14. Milking traits have been favored for a long time in Norwegian Red, and several QTLs of these trait are located at BTA14 [17], which may explain several fixed haplotypes at this chromosome.

BTA 12 revealed a gap of the available markers, restricting any ROH to be detected across this segment, also observed by Sölkner et al. [20]. Lack of SNP over large areas reduces the precision of ROH detections, and efforts should be done to find SNP markers at these deserts in order to map genetics in these areas as well.

Further insight could be obtained by refining findings obtained in this study. At the relevant segments, haplotypes need to be identified and followed over generations to examine which that are actually preferred through the selection process.

Conclusions

Ongoing selection signatures can be identified without using any phenotypic data by regressing the state of being in a ROH on time. Further insight can be obtained by visual inspection of distribution of ROH over time, allowing to discriminate between ongoing and historical sweeps.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors designed the study, interpreted the findings and revised the manuscript. BH ran the calculations. SAB designed the scripts and functions in R for illustrating ROH over time. BH, JAW, DIV and GK analyzed the results. BH drafted the manuscript. JAW, SAB, DIV and GK co-wrote the manuscript.

Acknowledgments

We would like to thank the Norwegian University of Life Sciences for funding this project. Geno, especially Morten Svendsen and Trygve Roger Solberg are acknowledged for sharing pedigree files and genotyping data. Thank you also to Trygve Flathen for solving programming issues when needed.

References

1. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, Mass.: Sinauer Associates; 1998.

2. Wright S. Coefficients of Inbreeding and Relationship. The American Naturalist. 1922;56(645):330-8. doi:10.2307/2456273.

3. VanRaden PM, Olson KM, Wiggans GR, Cole JB, Tooker ME. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science. 2011;94(11):5673-82. doi:10.3168/jds.2011-4500.

4. Saura M, Fernandez A, Rodriguez MC, Toro MA, Barragan C, Fernandez AI et al. Genome-Wide Estimates of Coancestry and Inbreeding in a Closed Herd of Ancient Iberian Pigs. Plos One. 2013;8(10):7. doi:10.1371/journal.pone.0078314.

5. Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection requires genomic control of inbreeding. Genet Sel Evol. 2012;44:10. doi:10.1186/1297-9686-44-27.

6. Hillestad B, Woolliams JA, Meuwissen T, Våge DI, Klemetsdal G. Estimating rate of inbreeding and effective population size using genomic data in Norwegian Red Submitted to GSE. 2015.

7. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974;23(01):23-35. doi:doi:10.1017/S0016672300014634.

8. Kim ES, Cole JB, Huson H, Wiggans GR, Van Tassell CP, Crooker BA et al. Effect of Artificial Selection on Runs of Homozygosity in US Holstein Cattle. Plos One. 2013;8(11). doi:10.1371/journal.pone.0080813.

9. Boichard D. Pedig: a fortran package for pedigree analysis suited to larger populations. 7th World Congress on Genetics Applied to Livestock Production; 19th-23rd of August; Montepellier2002. p. paper 28-13.

10. Maccluer JW, Boyce AJ, Dyke B, Weitkamp LR, Pfennig DW, Parsons CJ. Inbreeding and pedigree structure in standardbred horses. Journal of Heredity. 1983;74(6):394-9.

11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 812007.

12. Hillestad B, Woolliams JA, Boison SA, Grove H, Meuwissen T, Våge DI et al. Detection of runs of homozygosity in Norwegian Red: Density, critera and genotyping quality control. Submitted to GSE. 2015.

13. Cases S, Smith SJ, Zheng YW, Myers HM, Lear SR, Sande E et al. Identification of a gene encoding an acyl CoA:diacylglycerol acyltransferase, a key enzyme in triacylglycerol synthesis. Proc Natl Acad Sci U S A. 1998;95(22):13018-23.

14. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P et al. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research. 2002;12(2):222-31. doi:10.1101/gr.224202.

15. Gouveia JJD, da Silva M, Paiva SR, de Oliveira SMP. Identification of selection signatures in livestock species. Genet Mol Biol. 2014;37(2):330-42.

16. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013;28(11):659-69. doi:10.1016/j.tree.2013.08.003.

17. Wibowo TA, Gaskins CT, Newberry RC, Thorgaard GH, Michal JJ, Jiang Z. Genome assembly anchored QTL map of bovine chromosome 14. International journal of biological sciences. 2008;4(6):406-14. doi:citeulike-article-id:4363174.

18. Karlengen IJ, Harstad OM, Taugbol O, Berget I, Aastveit AH, Vage DI. The effect of excess cobalt on milk fatty acid profiles and transcriptional regulation of SCD, FASN, DGAT1 and DGAT2 in the mammary gland of lactating dairy cows. Journal of animal physiology and animal nutrition. 2012;96(6):1065-73. doi:10.1111/j.1439-0396.2011.01221.x.

19. Ashwell MS, Heyen DW, Sonstegard TS, Van Tassell CP, Da Y, VanRaden PM et al. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. J Dairy Sci. 2004;87(2):468-75. doi:10.3168/jds.S0022-0302(04)73186-0.

20. Sölkner J, Ferenčaković M, Karimi1 Z, Perez O‘Brien AM, Mészáros G, Eaglen S et al. Extremely Non-uniform: Patterns of Runs of Homozygosity in Bovine Populations. 10th World Congress of Genetics Applied to Livestock Production; Vancouver, Canada: American Society of Animal Science; 2014.

Table 1: Chromosomal rate of inbreeding over time

Chromosomal rate of inbreeding (ΔF) and corresponding effective population size (Ne) from ln(1-F_{Homj}) regressed on complete generation equivalence (CGE) in 375 Norwegian Red bulls, born between 1975 and 2004, genotyped with the Illumina 777K HD-panel. F_{Homj} are individual inbreeding coefficients utilized from observed homozygosity.

[1] ΔF and standard errors are scaled by 10^3.

[2] Chromosomes with *-marked p-values had nominal significance, while **-marked p-values referred to a Bonferroni significance under 0.05/29=0.0017.

BTA	ΔF[1]	Se[1]	Ne	p-value[2]
1	2.3	3.4	218.0	0.501
2	5.8	3.4	86.5	0.089
3	5.6	3.4	89.7	0.102
4	2.9	5.1	171.8	0.570
5	19.6	4.7	25.4	**0.000
6	16.4	5.2	30.5	*0.002
7	5.2	3.4	95.5	0.119
8	9.4	4.9	53.4	0.055
9	12.6	5.0	39.8	*0.012
10	1.8	5.2	274.4	0.724
11	13.0	4.4	38.4	*0.003
12	5.6	3.7	89.7	0.127
13	8.4	5.2	59.6	0.103
14	14.6	4.3	34.2	**0.001
15	12.8	5.3	39.1	*0.015
16	10.7	5.2	46.7	*0.039
17	7.6	4.6	66.1	0.103
18	9.2	5.1	54.1	0.071
19	6.8	4.0	73.4	0.086
20	14.9	5.7	33.5	*0.009

21	10.7	4.7	46.6	*0.023
22	1.2	6.9	418.0	0.862
23	13.7	6.8	36.6	*0.044
24	22.1	5.0	22.6	**0.000
25	10.2	5.6	48.9	0.066
26	1.8	5.8	280.8	0.760
27	10.7	6.2	46.9	0.085
28	11.2	6.0	44.7	0.061
29	inf	-	-	-

Table 2: Average numbers of ROH detection

Basic statistics of runs of homozygosity (ROH) detected in 381 Norwegian Red bulls, born between 1971 and 2004, genotyped with an Illumina HD-panel (708K).

Minimum length of ROH	0.5 Mb	2 Mb
Total # of segments	47,437	10,308
Mean length (Kb)	1,839	5,440
Standard deviation of length (Kb)	2,854	4,525
Median length (Kb)	824	3,884
Longest ROH (Kb)	58,724	58,724
Mean # of segments pr animal	125	27
Minimum # of segments pr animal	72	1
Maximum # of segments pr animal	185	56

Table 3: Average numbers of ROH detection at segments with high rate of inbreeding

Basic statistics of runs of homozygosity (ROH) in segments with a significantly increased frequency (-log(p) > 4) of ROH over time obtained in 381 Norwegian Red bull, born between 1971 and 2004, genotyped with an Illumina HD-panel (708K). Minimum length of ROH was set to 0.5 Mb.

BTA	Segment (Mb)	Mean length (Kb)	Median length (Kb)	Maximum length (Kb)	# ROH detected
5	70-95	1,703	711	32,508	576
6	45-65	2,445	732	58,724	539
12	10-20	2,068	1,131	36,773	186
24	10-20	2,162	974	16,347	123

Figure 1: Positional F-values from ROH in Norwegian Red

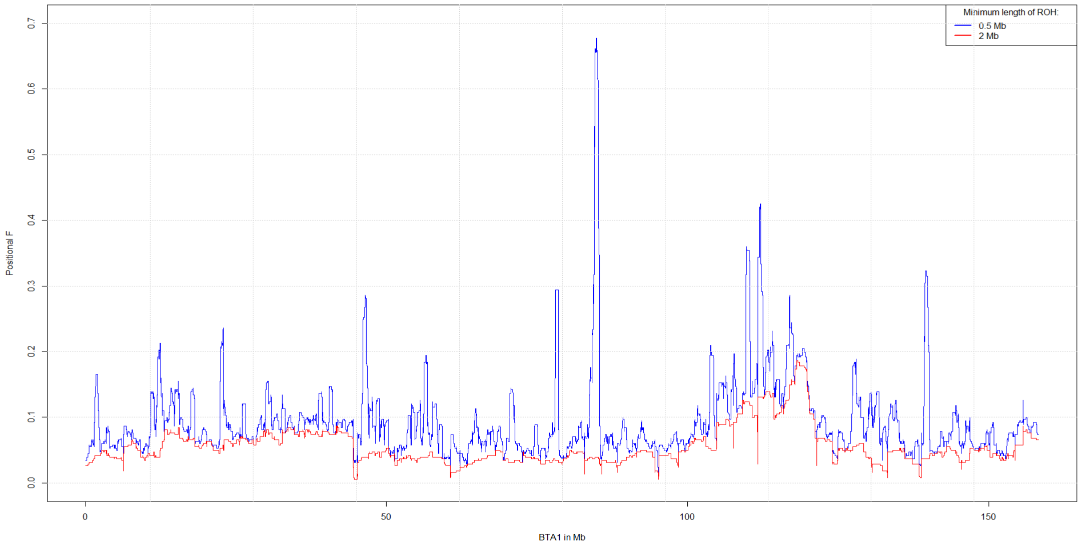
Graphs illustrating average positional inbreeding coefficients (F), from whether a SNP is within a runs of homozygosity (ROH) or not in BTA 1, 5, 7, 14 and 22, based on ROH with varying minimum length in 381 Norwegian Red bulls, born between 1971 and 2004, genotyped with an HD-panel.

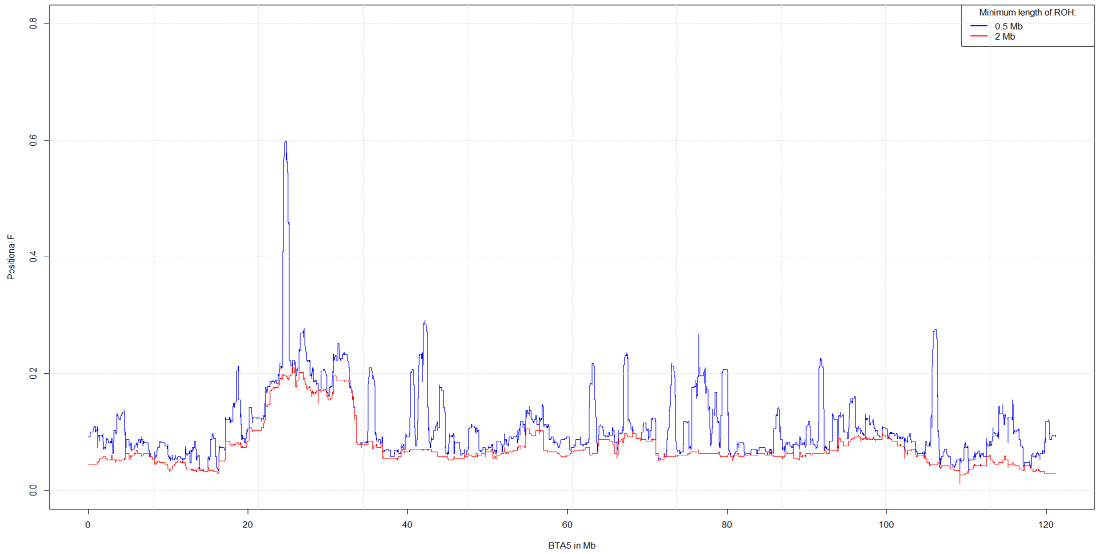
Figure 2: The slope of change of status at the locus Norwegian Red

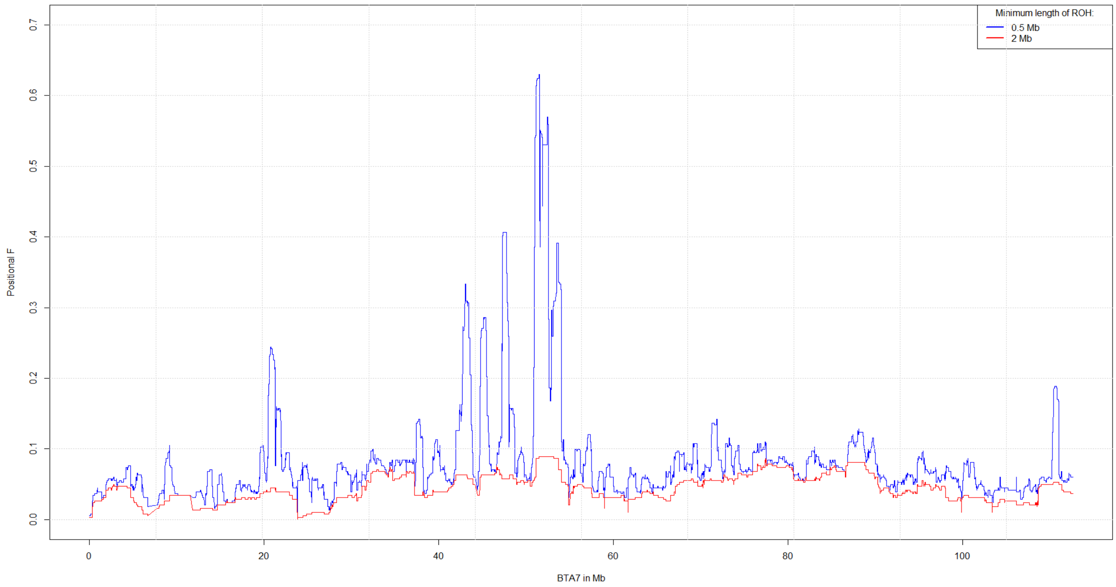
The slope of change of status at the locus per generation at BTA 5, 6, 12 and 24; whether a SNP is within a run of homozygosity (ROH) or not estimated by logistic regression in 381 Norwegian Red, born between 1971 and 2004, genotyped with an Illumina HD-panel. The black curve is the slope of a logistic regression done on each SNP whether or not is was within a ROH regressed on CGE. The red curve is the –log(p)-value of the regression.

Figure 3: ROH-plot over time in Norwegian Red

Distribution of runs of homozygosity (ROH) per animal on BTA 5, 6, 12, 14 and 24, in 381 Norwegian Red bulls, born between 1971 and 2004, genotyped the Illumina HD-panel. The animals are sorted on year of birth and ID-numbers, where the oldest animals are placed in the bottom of the plot and the youngest animals on the top. Ongoing selective sweeps are visible at BTA 5, 6, 12 and 24. Potential historical sweeps appears in all 5 chromosomes, but BTA 14 show complete fixation as what the product of a historical sweep actually is.


Minimum length of ROH:
0.5 Mb
2 Mb
Positional F
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
50
100
150
BTA1 in Mb


Minimum length of ROH:
0.5 Mb
2 Mb
Positional F
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
120
BTA5 in Mb


Minimum length of ROH:
0.5 Mb
2 Mb
Positional F
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
20
40
60
80
100
BTA7 in Mb

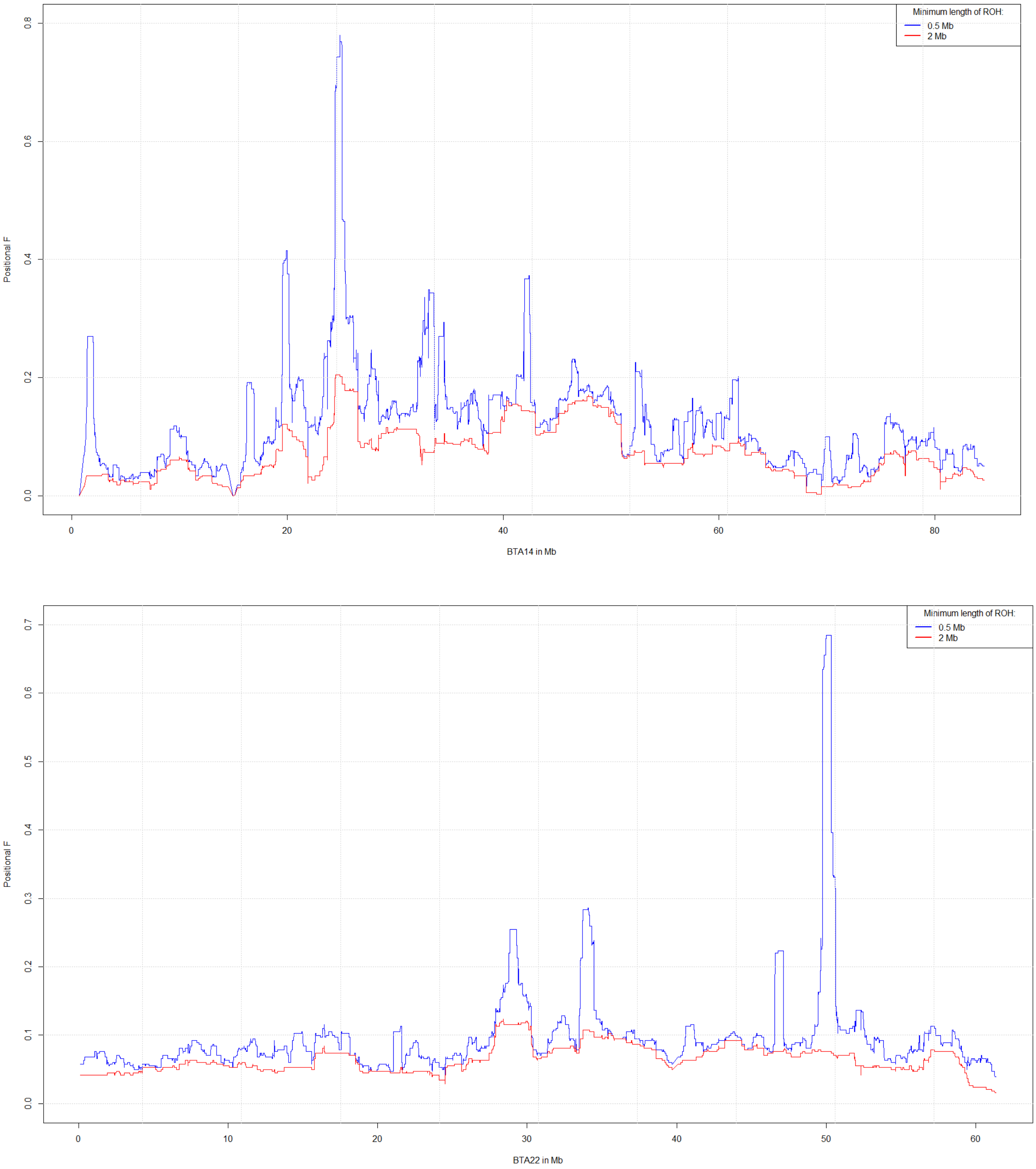
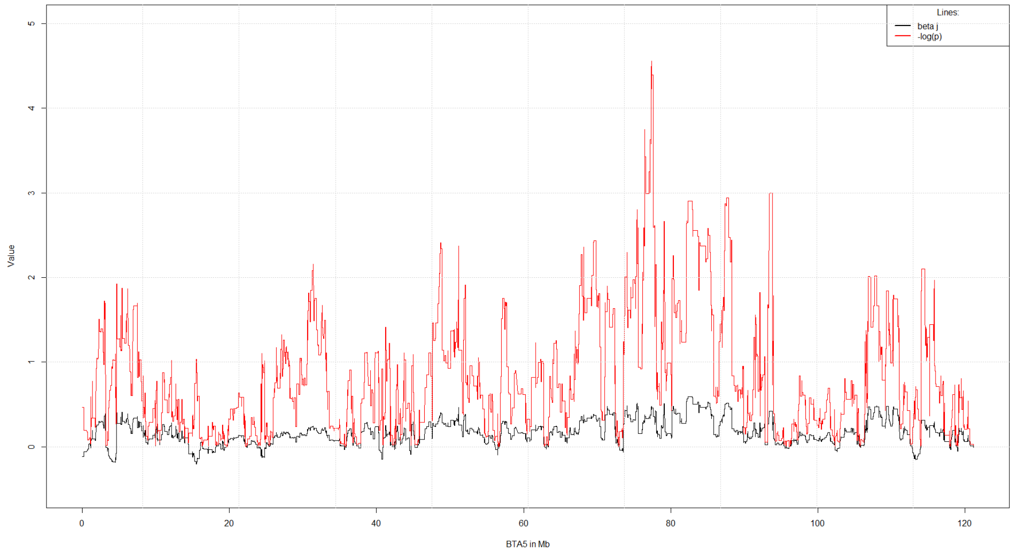
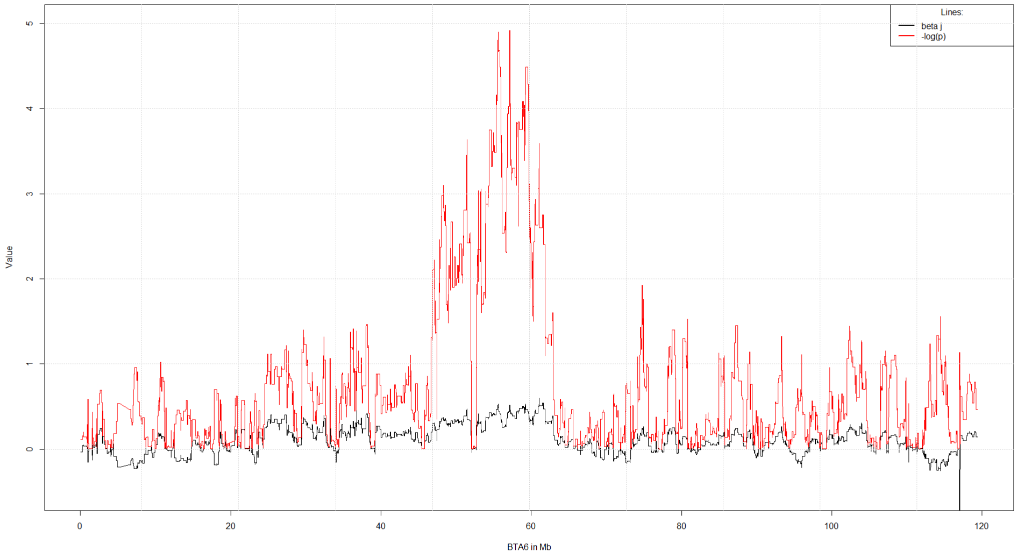
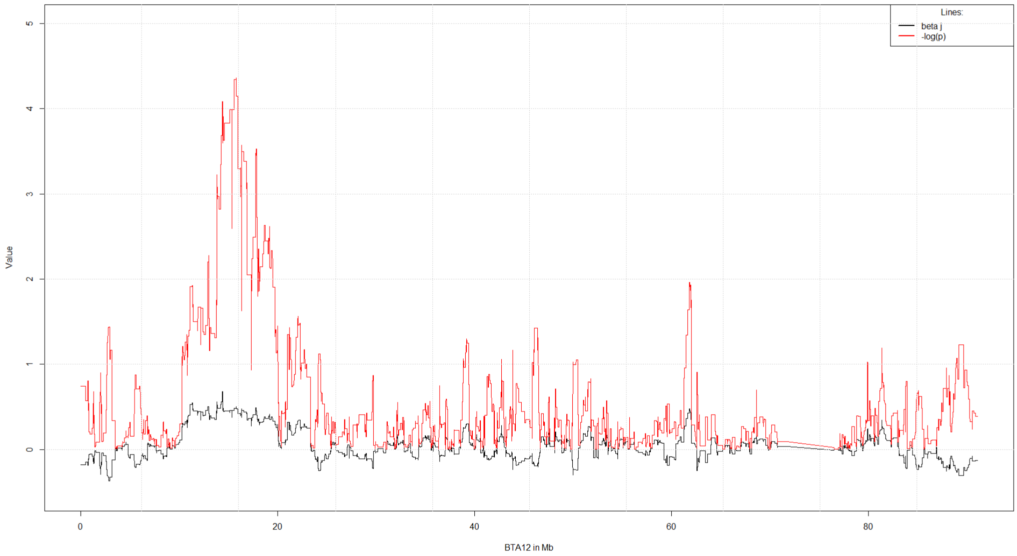

Minimum length of ROH:
0.5 Mb
2 Mb
Positional F
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
BTA14 in Mb
Minimum length of ROH:
0.5 Mb
2 Mb
Positional F
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50
60
BTA22 in Mb

Figure 1


Lines:
beta j
-log(p)
Value
0
1
2
3
4
5
0
20
40
60
80
100
120
BTA5 in Mb


Lines:
beta j
-log(p)
Value
0
1
2
3
4
5
0
20
40
60
80
100
120
BTA6 in Mb


Lines:
beta j
-log(p)
Value
0
1
2
3
4
5
0
20
40
60
80
BTA12 in Mb

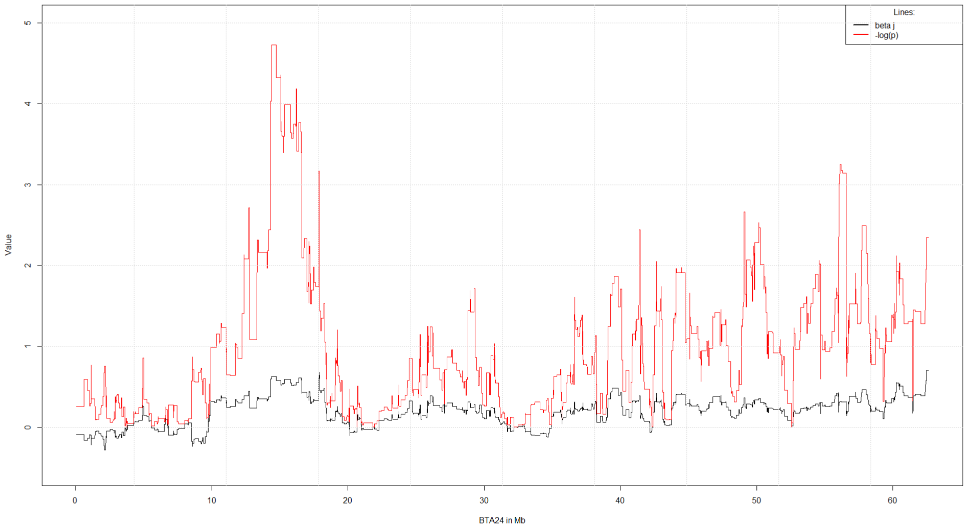
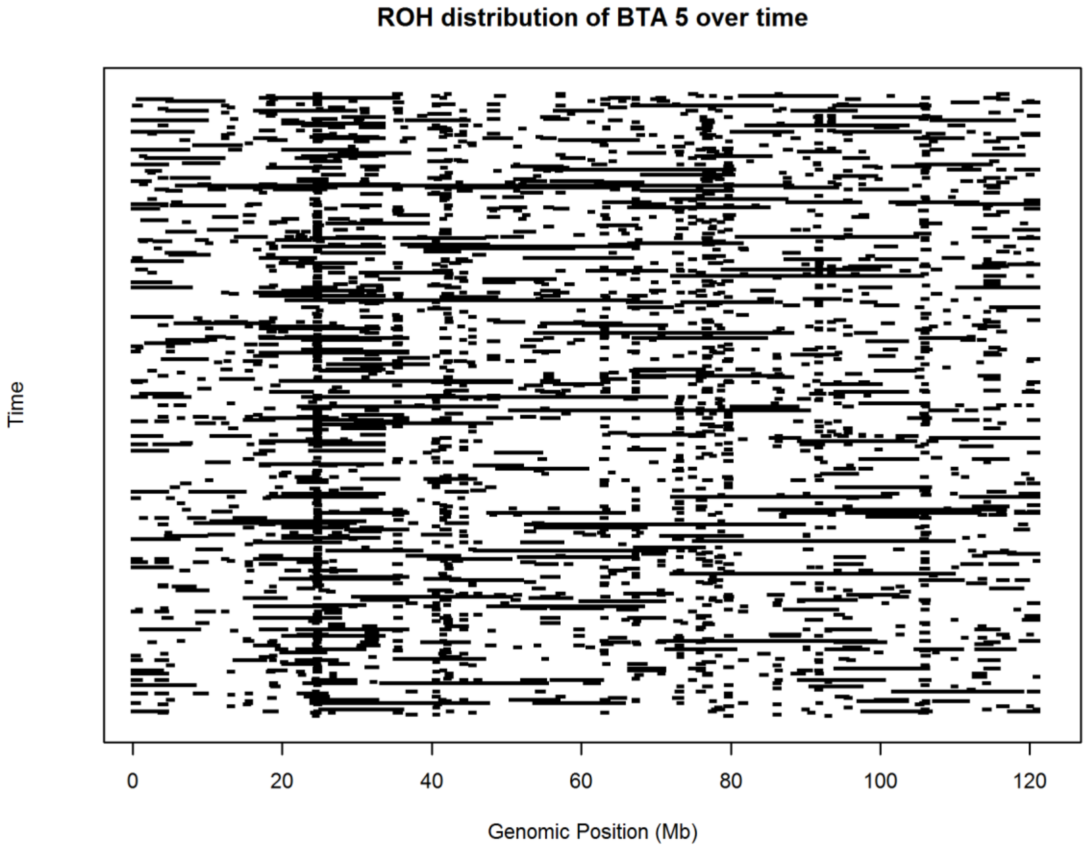
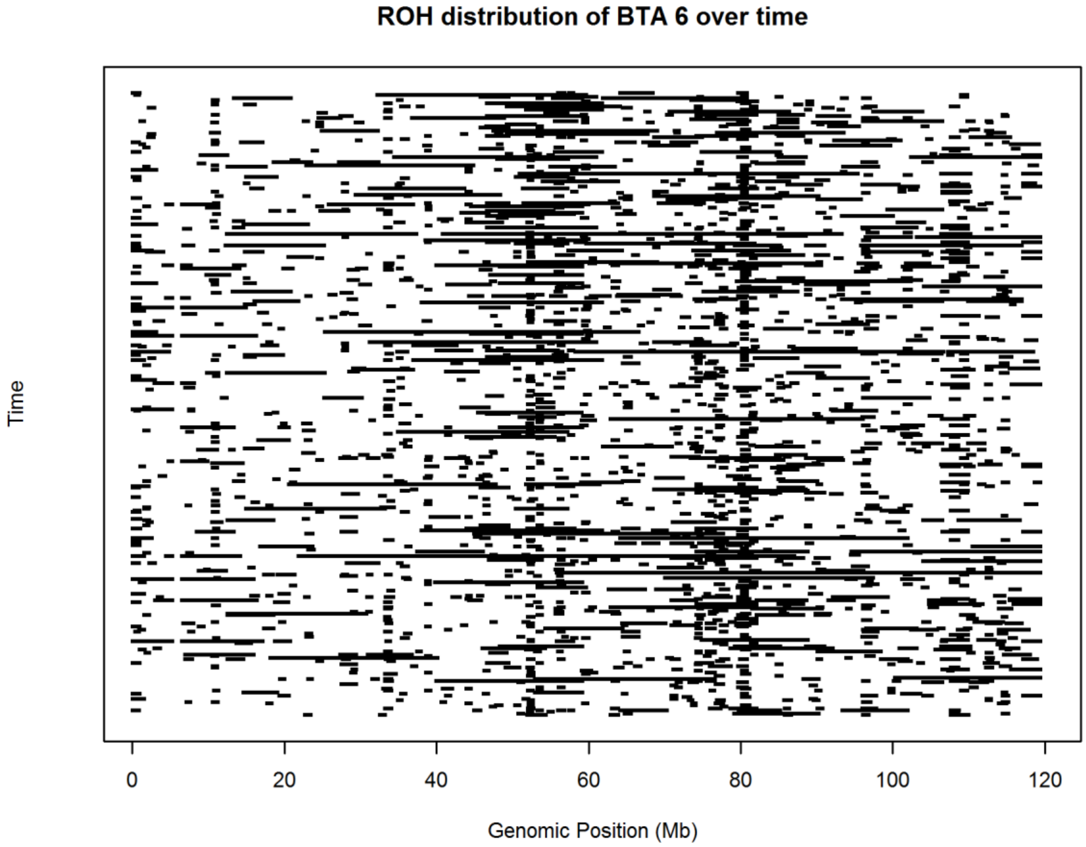
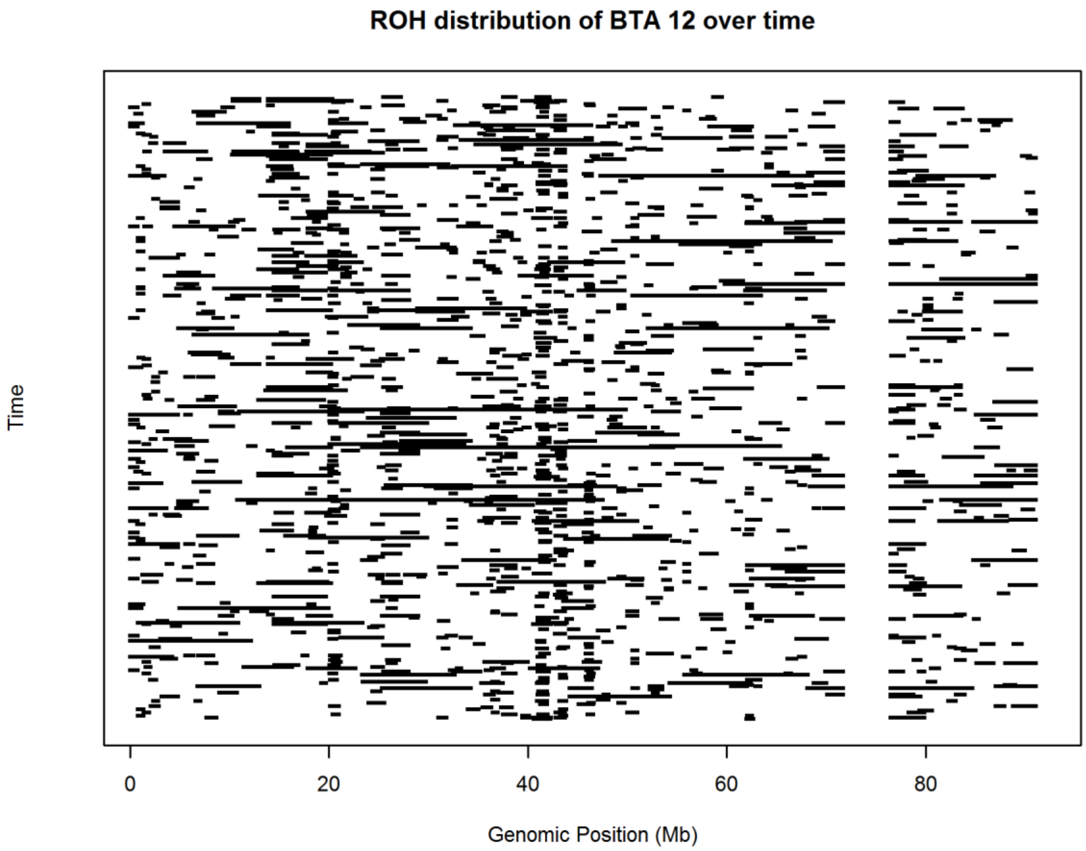


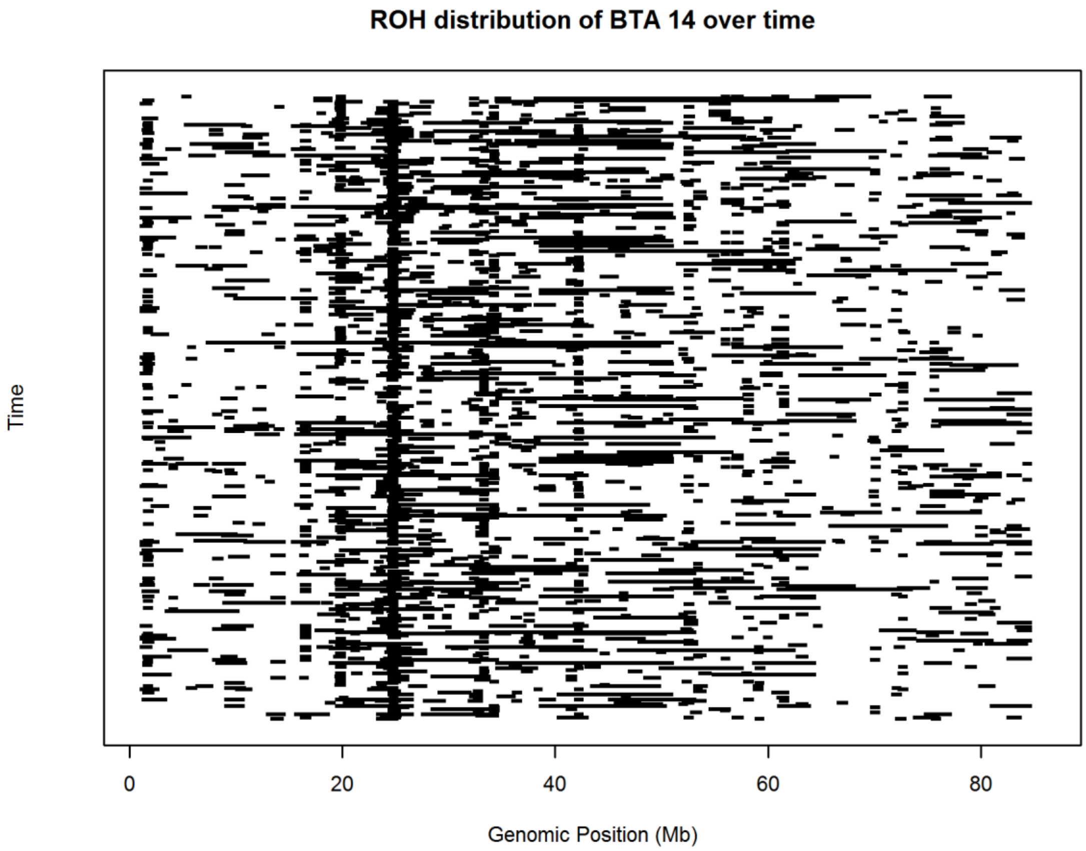
Figure 2


ROH distribution of BTA 5 over time
Time
0
20
40
60
80
100
120
Genomic Position (Mb)


ROH distribution of BTA 6 over time
Time
0
20
40
60
80
100
120
Genomic Position (Mb)

419


ROH distribution of BTA 12 over time
Time
0
20
40
60
80
Genomic Position (Mb)


ROH distribution of BTA 14 over time
Time
0
20
40
60
80
Genomic Position (Mb)

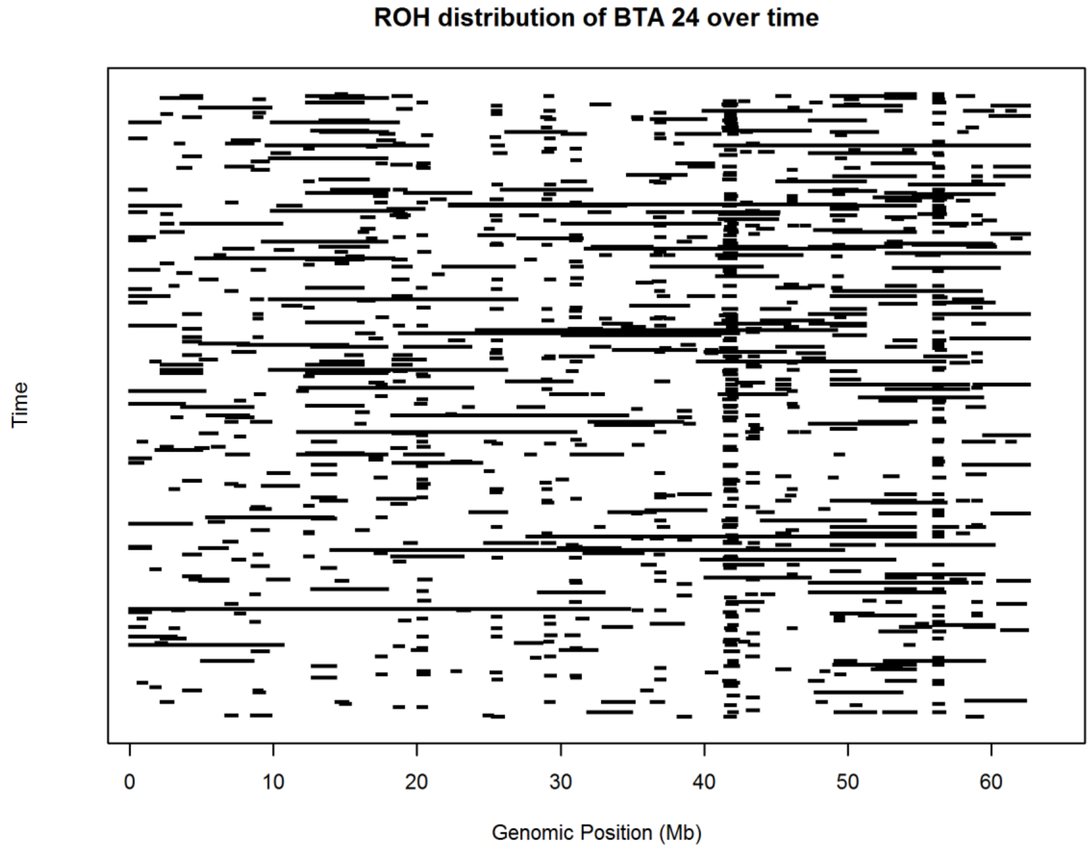

ROH distribution of BTA 24 over time
Time
Genomic Position (Mb)
0
10
20
30
40
50
60

Figure 3