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

Borghild Hillestad[1], John Arthur Woolliams[2,3], Solomon Antwi Boison[4], Harald Grove[5], Theo Meuwissen[2], Dag Inge Våge[2], Gunnar Klemetsdal[2]

[1]SalmoBreed AS, Sandviksboder 3A, N-5035 Bergen, Norway

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

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

[4]Nofima AS, Osloveien 1, N-1430 Ås, Norway

[5]Department of Research and Development, Mahidol University, 999 Phuttamonthon 4 Road, Salaya, Nakhon Pathom 73170 Thailand

Borghild Hillestad borghildhillestad@gmail.com

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

Solomon Antwi Boison soloboan@yahoo.com

Harald Grove harald.gro@mahidol.ac.th

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

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

Introduction

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 (Broman & Weber, 1999). Over its length, the frequency of homozygosity depends on the history and the management of the population. The use of molecular markers in 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. 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 is that, it has the possibility to distinguish between recent and ancient inbreeding (Hayes et al., 2003). A homozygous segment originating from a more recent ancestor is expected to be longer as there have been fewer opportunities for recombination to reduce its length. 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 (McQuillan et al., 2008).

However, these simple ideas have 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 (Bolormaa et al., 2010, Nishimura et al., 2012, Kim et al., 2013, Ferenčaković et al., 2013a). 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 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 (**GS**) or genetic relationship matrix (**G**) (Meuwissen et al., 2001). In the context of GS, it is common to delete SNP with MAF as high as 0.05 (Cole et al., 2009). Other studies like Keller et al. (2011) 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.

Another important factor is the density of the SNP chip used in ROH detection (Howrigan et al., 2011; Purfield et al., 2012; Ferenčaković et al., 2013b). Ferenčaković et al. (2013b) demonstrated that, when detecting ROH segments that are < 4 Mb, the use of the Illumina Bovine 50K SNP chip (the SNP chip commonly used in genomic evaluation in cattle populations) is not appropriate. They observed that, with the 50K SNP chip, the detected ROHs with length < 4 Mb were mostly artefact which led to an overestimation of F_{ROH} compared to the Illumina HD Bovine SNP chip, that keeps a SNP density of 777K. Although HD SNP chips have not been widely used as the default genotyping array due to it cost, there is currently an increasing tendency to use a slightly denser SNP array for genomic evaluation in cattle. The reasons for using a denser SNP array varies from the possibility of including causal variants detected with the BovineHD or sequence information, and availability of relatively cheaper and more informative SNP chips (GeneSeek [Neogen Corp., Lexington, KY] vs. Illumina [Illumina Inc., San Diego, CA]), among others. For example, there is a gradual shift from the 50K SNP chip to the 77K/84K SNP array by the Council on Dairy Cattle Breeding (Bowie, MD) in the United States (Wiggans et al., 2016). There is therefore the potential of using different SNP

densities (not only the Bovine 50K and HD) in the detection of ROHs, and these need to be studied.

In addition to the impact of SNP density on detecting ROHs, there is lack of uniformity in criteria used for the detection of a ROH segment. This lack of uniformity is due to the complexities in defining: i) the size (the number of markers or length of segment) of the sliding window; ii) the minimum ROH length (either in number of markers or segment length); iii) the number of markers allowed to be missing within a sliding window and iv) the number of heterozygotes allowed (Purfield et al., 2012; Ferenčaković et al., 2013b; Sölkner et al., 2014; Marras et al., 2015; Mészáros et al., 2015). The lack of standards in the criteria used for ROH detection could be attributed to: a) difficulties in applying ROH detection standards across species (e.g. standards from human genetic studies cannot directly be applied to cattle or chicken populations due to difference in effective population size), or b) differences in pattern of genotyping errors, quality of genotypes, or allele frequency distribution for different SNP panels. This therefore restricts the direct adoption of ROH detection criteria from different authors. For example, after a careful study of different ROH criteria for detection, Ferenčaković et al. (2013b) concluded that, the number of heterozygous SNPs allowed within a ROH segment, should be determined separately for each ROH length of interest and for each SNP density. Since the criteria to define ROH for each SNP density 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.

116

117 Materials and Methods

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

119 The impact of SNP density on the detection of ROH was examined in 384 Norwegian Red bulls
120 genotyped with the Illumina HD panel. The panel contains 777,962 SNP-markers, covering 2.51
121 Gb of the 3 Gb large genome, although not all these SNP-markers will be polymorphic in the
122 Norwegian Red. After genotyping, the marker data passed through several stages of quality
123 controls, or genotype editing, to exclude markers on sex-linked chromosomes, call rate per SNP
124 < 90 % (individual SNP score missing if GenCall score < 0.7) and deviation from Hardy-
125 Weinberg ($P < 10^{-6}$) (Table 1). Three animals were deleted for having genotypes for fewer than
126 95 % of loci. This resulted in the retention of 707,609 SNP, which will be denoted the 708K set.

127

128 The 708K set was sequentially pruned to give further nine subsets of data. The pruning was done
129 to test the effect of SNP density on the size of detectable ROHs. Recommendation from the
130 results of testing different SNP densities is especially useful in the cattle breeding industry where
131 different SNP arrays are used for genomic evaluation and invariably ROH detection (Neves et
132 al., 2014; Haile-Mariam et al., 2015; Wiggans et al., 2016). The first pruning removed every
133 fourth SNP, by physical order, from the 708K set to obtain a subset of 530,706 SNP (denoted
134 531K set). This procedure was repeated by removing every fourth SNP from the 531K set, to
135 obtain a 398K set, and a further seven times to give the smallest subset (53K set). All densities
136 achieved are shown in Table 2.

137

For each of these sets, ROH were identified with PLINK 1.07 (Purcell et al., 2007). PLINK takes a window of 5,000 Kb and slides it across the genome, determining homozygosity at each window. The identifications of ROH in PLINK 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 a broad specter of criterion parameters were tested in advance. Since the number of SNPs analyzed per sliding window increased with SNP density, the parameter settings chosen were changed accordingly, and the settings are shown in Table 3.

Detection of ROH when altering ROH criteria

When searching for ROHs, it has been common to allow one heterozygote SNP per window, because they are assumed to be genotyping errors. Normally, you would not expect to find heterozygote SNP in a window that only contains homozygote SNPs, but this step may provide false ROHs as the density on arrays over time are increasing and the genotyping technology is improving. Therefore, to test the effect of allowing one heterozygote SNP per window another subset ($708K_{Alt1}$) was generated that did not allow for any heterozygote SNP per window (Table 3). Further, the effect of applying ROH criteria used for lower SNP density sets was examined by generating three datasets; $708K_{Alt2}$, $708K_{Alt3}$ and $708K_{Alt4}$, that used the same criteria applied to the 53-94K, 126K and 168-299K SNP densities, respectively. In addition to not allowing a

heterozygous SNP within a ROH for the 708K SNP density ($708K_{Alt1}$), the number of SNPs allowed to be missing in a ROH was reduced from 3 to 1 SNP ($708K_{Alt5}$).

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 ($611K_{MAF}$). A further subset was obtained by removing SNP with MAF < 0.02; resulting in a loss of an additional 2 % of SNP and a total number of 597,454 SNP ($597K_{MAF}$) (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, $611K_{MAF}$ and $597K_{MAF}$ were investigated and classified according to chromosomes.

Heterozygosity on a chromosomal level

To search for signs of selection, heterozygosity was estimated at a chromosomal level. For the 708K set, average rate of heterozygosity (**Het**) was estimated based on the following equation:

$$\mathrm{Het} = \mathrm{O}(\mathrm{Het}) / \mathrm{N}(\mathrm{NM}) \quad (1)$$

where O(Het) is observed heterozygosity and N(NM) is defined as the number of non-missing genotypes.

Results

Variation in SNP densities and ROH criteria

Minimum number of homozygous SNP/Kb. With a minimum threshold set both in Kb and in number of SNP, this is reflected in the missing pattern of Table 4, e.g. ROH segments 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 $708K_{Alt2}$ sets; 126K and $708K_{Alt3}$ sets; 224K, 299K and $708K_{Alt4}$ sets and the 398K, 531K and 708K sets, where in each of these groups all criteria was the same except for the density that was altered (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 false positives ROH segments that may have been identified with the lower SNP density (Figure 1a). This is because an increasing density of markers within a ROH will allow for detection of heterozygote markers not present on the lower density marker panel. For the first group (53K, 71K, 94K and $708K_{Alt2}$ sets) the lengths of ROH seemed to be redistributed when density was changed (Table 4), because as SNP density increased, longer ROH were split into shorter segments, which reduced the total length of ROH.

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 of 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 (Figure 1b). 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 allowance of heterozygote SNP in a sliding window will increase the number of false positive ROHs, and is therefore not recommended.

The average heterozygosity frequency within all ROHs at the 708K set was 1.1%. In this density the minimum length of ROH was set to 0.5 Mb, and the frequency was higher in the 0.5-1 Mb group (1.4%). In addition, the total number or called ROH in this group was 49,965 compared to 70,148 overall. Given that it for this density is estimated to be on average 1,179.3 SNPs on average per 5 Mb sliding window (Table 3) and the we have allowed one heterozygote SNP per sliding window, the frequency of heterozygosity within a run should be closer to $8x10^{-4}$. When considering the 4-8 Mb ROH group in this dataset, the frequency of heterozygosity was in total accordance with this estimate, and had a heterozygosity frequency of $8x10^{-4}$.

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 2).

Missing SNP. The effect of allowing three missing SNP per window vs only one missing SNP was examined (Table 4: $708K_{Alt1}$ vs $708K_{Alt5}$). 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.

MAF. By removing low MAF SNP from the data, the amount of long ROH increased and the amount of short ROHs decreased (Figure 1c). 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 ROHs 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 SNPs 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 SNPs: 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).

All ROH results presented in this study was found using PLINK 1.07, but as an extra control, we also ran the dataset by SNP & Variation Suite 8.8.1 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). The outcome from SVS analysis was highly similar to the outcome from PLINK 1.07, and was therefore not further looked into (results not presented).

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 (Ferenčaković et al., 2013b). In this study we altered on common variables and constraints within SNP density, genotyping quality controls and criteria to detect ROH when using PLINK 1.07, where several factors rather gained than removed error.

As the results showed, a redistribution of ROH occurred as the SNP density increased. Naturally as the SNP density increases, both homozygote and heterozygote SNPs will occur in the newly added SNPs, also in stretches of ROHs. This will cause a breakdown of ROHs and an increase of short ROHs will arise together with a decrease of long ROHs. Therefore, 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 contribute 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

297 ROH of 0.5 Mb was defined in accordance with Purfield et al. (2012) and their study of multiple
298 cattle breeds (Angus, Belgian Blue, Charolais, Friesian, Hereford, Holstein, Holstein-Friesian
299 crosses, Limousin and Simmental), although there are several strategies for the minimum length
300 threshold. Ferenčaković et al. (2013a) chose 1 Mb as the minimum length when studying Brown
301 Swiss, Pinzgauer, Tyrol Grey cattle to avoid ROHs that were more likely to arise due to
302 population linkage disequilibrium (LD) rather than due to inheritance. Sodeland et al. (2011)
303 showed low LD levels at 0.5 Mb ($r^2 < 0.1$) in a historical analysis of Norwegian Red, which
304 strengthens our confidence in not calling ROHs aroused due to LD by setting the minimum
305 length of 0.5 Mb. There have been speculations whether or not it would be appropriate to raise
306 the minimum length of ROH in order to capture recent inbreeding and avoid ancient inbreeding
307 that no longer concerns the population, which is why the minimum length has been raised in
308 some studies (Rodriguez-Ramilo et al., 2015, Gómez-Romano et al., 2014). When inbreeding
309 was measured by ROH, all homozygosity that where not defined to be within a ROH was
310 rejected and assumed not to be IBD. Because we do not know if this assumption is correct, and
311 because some of the approved ROH also may not be IBD, we should be careful about removing
312 even more homozygosity by raising the threshold of minimum length. Precision is increased by
313 keeping as much information on homozygote SNP as possible.
314
315 Although changing the threshold in certain criteria set to define ROH did not influence on the
316 detection of ROH in most cases, two main criteria need to be commented: (i) First, to account for
317 genotyping errors, the ROH criterion allowed for one heterozygous SNP in a homozygous
318 segment within a window. This criterion created many short false positive ROH and should be
319 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 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. Two additional criteria that were tested (result not shown) and which did not have a strong effect on the number and size of ROHs detected were (iii) the average Kb per SNP and iv) maximum gaps between markers in a ROH. This was because, the average distance between markers on the HD panel is < 5 Kb, thus imposing a restriction of 50 Kb does not affect ROH detection. Furthermore, very few gaps between SNP will be long, especially when low MAF SNP were included and not pruned away, giving small differences in results when different gap lengths were studied. Overall, 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 genotyping error could be controlled by both a GC score threshold (Illumina, 2005) 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. Using markers with MAF > 0.01 and > 0.02 reduced the number of SNP by 14 % and 16 %, respectively, which might have led to the 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, incorrectly defined polymorphic SNP that contributed inaccurately and little to genomic evaluation estimation have been removed (Edriss et al., 2013, Wiggans et al., 2009). Removal of low MAF SNP was also custom in earlier studies within ROH (Ferenčaković et al, 2013a, Howrigan et al., 2011, Edriss et al., 2013, Kirin et al., 2010, Silió et al., 2013), however, recent literature has been in support of including information on low MAF SNP when searching for ROH (Ferenčaković et al, 2013b). 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. The latter is affected by the criteria used for identifying ROHs, which generally include a minimum number of SNPs within a run, a maximum gap length between adjacent SNPs, and a minimum SNP density within a run.

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 (Ramey et al., 2013). 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 segregate in other populations. 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. For instance, BTA 14, that showed a large amount of ROH and a low Het-value, has earlier proven to contain several gene variants that influences economical important traits for both milk and beef cattle breeds (Wibowo et al., 2008). 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

Bolormaa, S., Pryce, J.E., Hayes, B.J., Goddard, M.E. (2010) Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science*. 93(8):3818-33.

Broman, K.W., Weber, J.L. (1999) Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. *Am. J. Hum. Genet.*, 65(6):1493-500.

409 Cole, J.B., VanRaden, P.M., O'Connell, J.R., Van Tassell, C.P., Sonstegard, T.S., Schnabel, R.D.
410 et al. (2009) Distribution and location of genetic effects for dairy traits. *Journal of Dairy Science.*
411 92(6):2931-46.

412 Edriss, V., Guldbrandtsen, B., Lund, M.S., Su, G. (2013) Effect of marker-data editing on the
413 accuracy of genomic prediction. *Journal of Animal Breeding and Genetics.* 130(2):128-35.

414 Ferenčaković, M., Hamzić, E., Gredler, B., Solberg, T.R., Klemetsdal, G., Curik, I. et al. (2013a)
415 Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected
416 cattle populations. *Journal of Animal Breeding and Genetics*. 130(4): 286-293.

417 Ferenčaković, M., Sölkner, J., Curik, I. (2013b) Estimating autozygosity from high-throughput
418 information: effects of SNP density and genotyping errors. *Genetics, selection, evolution : GSE.*
419 45(1):42-.

420 Gómez-Romano, F., Sölkner, J., Villanueva, B., Mészáros, G., Cara, M.A.R., O'Brien, A.M.P. et
421 al., (2014) Genomic estimates of inbreeding and coancestry in Austrian Brown Swiss cattle.
422 *WCGALP*; 2014; Vancouver, Canada.

423 Haile-Mariam, M., Pryce, J. E., Schrooten, C. and Hayes, B. J. (2015). Including overseas
424 performance information in genomic evaluations of Australian dairy cattle. *J. Dairy Sci.*
425 Available from: http://www.journalofdairyscience.org/article/S002203021500171X/fulltext

426 Hayes, B.J., Visscher, P.M., McPartlan, H.C., Goddard, M.E. (2003) Novel multilocus measure
427 of linkage disequilibrium to estimate past effective population size. *Genome Research*.
428 13(4):635-43.

429 Howrigan, D., Simonson, M., Keller, M. (2011) Detecting autozygosity through runs of
430 homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics.*
431 12(1):460.

432 Illumina. (2005) Illumina GenCall Data Analysis Software. *www.illumina.com.* 2005.
433 http://res.illumina.com/documents/products/technotes/technote_gencall_data_analysis_software.
434 pdf.

435 Keller, M., Visscher, P., Goddard, M. (2011) Quantification of inbreeding due to distant
436 ancestors and its detection using dense SNP data. *Genetics*. 12:460.

437 Kim, E.S., Cole, J.B., Huson, H., Wiggans, G.R., Van Tassell, C.P., Crooker, B.A. et al. (2013)
438 Effect of Artificial Selection on Runs of Homozygosity in US Holstein Cattle. *Plos One*. 8(11).

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

441 Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C., Ajmone-Marsan, P., Valentini, A., Williams,
442 J. L. and MacCiotta, N. P. P. (2015). Analysis of runs of homozygosity and their relationship
443 with inbreeding in five cattle breeds farmed in Italy. *Anim. Genet.* 46:110–121.

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

446 Mészáros, G., Boison, S. A., Perez O'Brien, A. M., Ferenčaković, M., Curik, I., Da Silva, M. V.
447 B., Utsunomiya, Y. T., Garcia, J. F. and Sölkner, J. (2015). Genomic analysis for managing
448 small and endangered populations: a case study in Tyrol Grey cattle. *Front. Genet.* 6:1–12.
449 Available from:

http://www.frontiersin.org/Livestock_Genomics/10.3389/fgene.2015.00173/abstract

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157(4):1819-29.

Neves, H. H., Carvalheiro, R., O`Brien, A. M., Utsunomiya, Y. T., do Carmo, A. S., Schenkel, F. S., Sölkner, J., McEwan, J. C., Van Tassell, C. P., Cole, J. B., da Silva, M. V., Queiroz, S. A., Sonstegard, T. S., and Garcia, J. F. (2014). Accuracy of genomic predictions in Bos indicus (Nellore) cattle. *Genet. Sel. Evol.* 46:17. Available from: http://www.gsejournal.org/content/46/1/17

Nishimura, S., Watanabe, T., Mizoshita, K., Tatsuda, K., Fujita, T., Watanabe, N. et al. (2012) Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. *Bmc Genetics*.13.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D. et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. [Software]. *American Journal of Human Genetics*, 812007.

Purfield, D. C., Berry, D. P., McParland, S. and Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genet.* 13:70. Available from: http://www.biomedcentral.com/1471-2156/13/70

Ramey, H.R., Decker, J.E., McKay, S.D., Rolf, M.M., Schnabel, R.D., Taylor, J.F. (2013) Detection of selective sweeps in cattle using genome-wide SNP data. *Bmc Genomics*. 2013;14.

Rodríguez-Ramilo, S.T., Fernández, J., Toro, M.A., Hernández, D., Villanueva, B. (2015) Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population. *PLoS One.* 10(4).

Silió, L., Rodríguez, M.C., Fernández, A., Barragán, C., Benítez, R., Óvilo, C. et al. (2013) Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. *Journal of Animal Breeding and Genetics.* 130(5):349-360.

SNP & Variation Suite ™ (Version 8.8.1) [Software]. Bozeman, MT: Golden Helix, Inc. Available from http://www.goldenhelix.com.

Sodeland, M., Kent, M., Hayes, B., Grove, H., and Lien, S (2011) Recent and historical recombination in the admixed Norwegian Red cattle breed. *BMC Genomics.* 12:33 http://www.biomedcentral.com/1471-2164/12/33.

Sölkner, J., Ferenčaković, M., Karimi, Z., O'Brien, A. M. P., Mészáros, G., Eaglen, S., Boison, S.A. and Curik, I. (2014). Extremely Non-uniform: Patterns of Runs of Homozygosity in Bovine Populations. *Proc. 10th WCGALP*, Vancouver, Canada.

Wibowo, T.A., Gaskins, C.T., Newberry, R.C., Thorgaard, G.H., Michal, J.J., Jiang, Z. (2008) Genome assembly anchored QTL map of bovine chromosome 14. *International journal of biological sciences.* 4(6):406-14.

Wiggans, G., Cooper, T., VanRaden, P., Van Tassell, C., Bickhart, D., and Sonstegard, T. (2016). Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. *J. Dairy Sci* 99:1–8. Available from: http://dx.doi.org/10.3168/jds.2015-10456

489 Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor,
490 J.F. et al. (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in
491 genomic evaluation of dairy cattle in the United States and Canada. *Journal of Dairy Science.*
492 92(7):3431-6.

493 **Table 1: Genotyping quality controls**

494 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: SNP densities used to detect ROH in Norwegian Red

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

This table shows the constraints that were set to detect ROH in Norwegian Red for datasets based on the following: i) Different SNP densities ranging from 53-708K after genotyping quality controls; ii) HD panels ($708K_{Alt1-5}$) where different constraints have been explored at the PLINK settings of ROH constraints and iii) HD panels with two different thresholds for MAF: One set where SNP with MAF < 0.01 were pruned ($611K_{MAF}$) and another at MAF < 0.02 ($597K_{MAF}$).

SNP density	SNP pr window (5,000 Kb)	Min. # homozygous SNP	Min. # homozygous Kb	# heterozygote SNP allowed per window	# missing SNP allowed per 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							
$708KAlt_1$	1,179.3	50	500	0	3	250	50
$708KAlt_2$	1,179.3	15	2,000	0	1	1,000	150
$708KAlt_3$	1,179.3	25	1,000	0	2	500	150
$708KAlt_4$	1,179.3	25	1,000	1	2	250	50
$708KAlt_5$	1,179.3	50	500	0	1	250	50
MAF sets							
$597K_{MAF}$	995.8	50	500	1	3	250	50
$611K_{MAF}$	1,018.1	50	500	1	3	250	50

505 **Table 4: Average number of detected ROH per animal**

506 Average number of ROH detected per individual, grouped into lengths of the segment in 381

507 Norwegian Red. Standard errors (SE) are listed in parentheses.

508

SNP density	0.5-1Mb	1-2Mb	2-4Mb	4-8Mb	8-16Mb	>16Mb	Total	Total >2Mb
Main density sets								
53K	-	-	9.8 (0.21)	8.0 (0.18)	4.0 (0.12)	1.4 (0.09)	23.2 (0.42)	23.2 (0.42)
71K	-	-	12.9 (0.24)	8.0 (0.18)	3.9 (0.12)	1.4 (0.09)	26.2 (0.45)	26.2 (0.45)
94K	-	-	13.1 (0.25)	8.0 (0.18)	3.9 (0.12)	1.4 (0.09)	26.4 (0.46)	26.4 (0.46)
126K	-	22.1 (0.26)	13.1 (0.25)	8.0 (0.18)	3.9 (0.12)	1.3 (0.09)	48.4 (0.57)	26.7 (0.46)
168K	-	36.2 (0.31)	14.0 (0.25)	8.0 (0.17)	3.9 (0.12)	1.5 (0.09)	63.6 (0.58)	27.4 (0.45)
224K	-	33.1 (0.31)	13.5 (0.25)	8.2 (0.18)	3.9 (0.12)	1.4 (0.09)	60.1 (0.59)	27.0 (0.46)
299K	-	30.4 (0.30)	13.6 (0.25)	8.2 (0.19)	3.9 (0.12)	1.3 (0.09)	57.4 (0.59)	27.0 (0.46)
398K	153.8 (0.67)	28.6 (0.28)	13.4 (0.25)	8.1 (0.18)	3.9 (0.12)	1.3 (0.09)	209.1 (0.80)	26.7 (0.46)
531K	142.4 (0.62)	27.4 (0.28)	13.4 (0.25)	8.0 (0.18)	3.9 (0.12)	1.3 (0.09)	196.4 (0.78)	26.6 (0.46)
708K	131.1 (0.61)	26.3 (0.29)	13.4 (0.25)	8.1 (0.18)	3.9 (0.12)	1.3 (0.09)	184.1 (0.79)	26.7 (0.46)
Variants of the HD-panel								
$708K_{Alt1}$	89.3 (0.51)	23.0 (0.31)	14.1 (0.27)	8.4 (0.20)	3.6 (0.12)	1.0 (0.08)	139.4 (0.83)	27.1 (0.50)
$708K_{Alt2}$	-	-	14.4 (0.29)	8.2 (0.20)	3.5 (0.12)	0.9 (0.08)	27.0 (0.51)	27.0 (0.51)
$708K_{Alt3}$	-	23.2 (0.31)	14.0 (0.28)	8.3 (0.19)	3.7 (0.12)	1.0 (0.09)	50.2 (0.66)	27.0 (0.50)
$708K_{Alt4}$	-	26.5 (0.30)	13.5 (0.26)	8.1 (0.19)	3.8 (0.12)	1.3 (0.09)	53.2 (0.61)	26.7 (0.47)
$708K_{Alt5}$	90.0 (0.58)	24.0 (0.39)	14.6 (0.29)	8.3 (0.20)	3.4 (0.12)	0.9 (0.08)	141.2 (1.00)	27.2 (0.52)
MAF sets								
$597K_{MAF}$	120.3 (0.59)	25.3 (0.28)	13.0 (0.25)	8.0 (0.18)	3.8 (0.12)	1.3 (0.09)	171.7 (0.79)	26.1 (0.46)
$611K_{MAF}$	121.9 (0.59)	25.5 (0.28)	13.0 (0.25)	8.0 (0.18)	3.8 (0.12)	1.3 (0.09)	173.5 (0.79)	26.1 (0.46)

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

510 Total loss of SNP per chromosome and short ROH (0.5-1Mb) by pruning for low MAF SNP and

511 average heterozygosity (Het) in 381 Norwegian Red genotyped with the 708K set.

BTA	Size of BTA in Mb *	Total SNP	Avg. # 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

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

Figure 1: Visualization of ROH segments identified for chromosome 5 using animals (n = 65) with the highest proportion of ROH. Each line represents one animal.

a) ROH identified with datasets of different densities; 53K and 708K: common to both (black), only in 53K (green) and only in 708K (red). Constraints are given in Table 3.

b) ROH identified with $708K_{Alt1}$ and 708K: common to both (black), only in $708K_{Alt1}$ (blue) and only in 708K (red). Both datasets with the same constraints (Table 3) with, respectively, one and no heterozygote allowed in a window.

c) ROH identified with $597K_{MAF}$ and 708K: common to both (black), only in $597K_{MAF}$ (blue) and only in 708K (red). Both datasets with the same constraints (Table 3) except for minor allele frequency (MAF) > 0.02 in $597K_{MAF}$.

Figure 2: 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 ($708K_{Alt1}$).

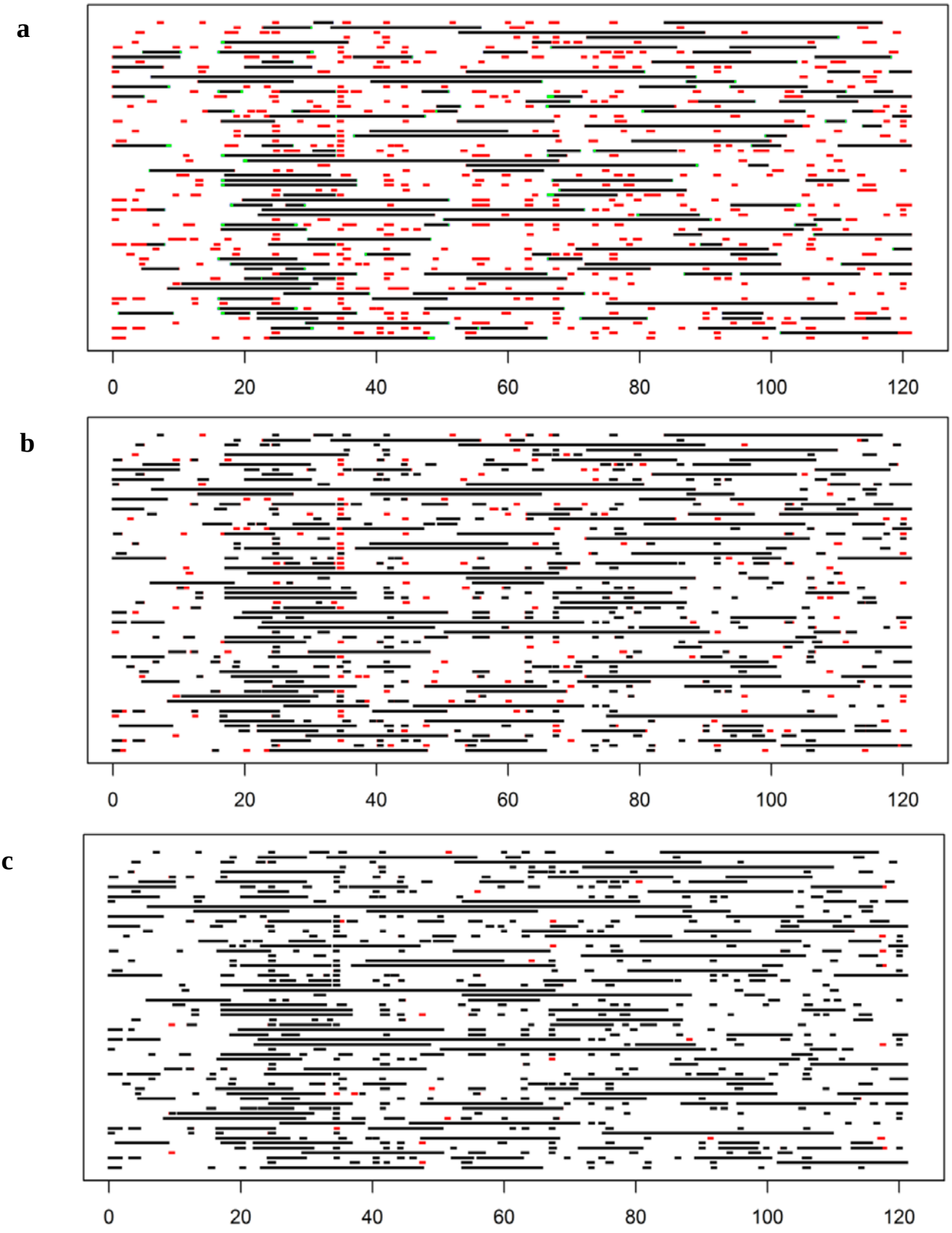
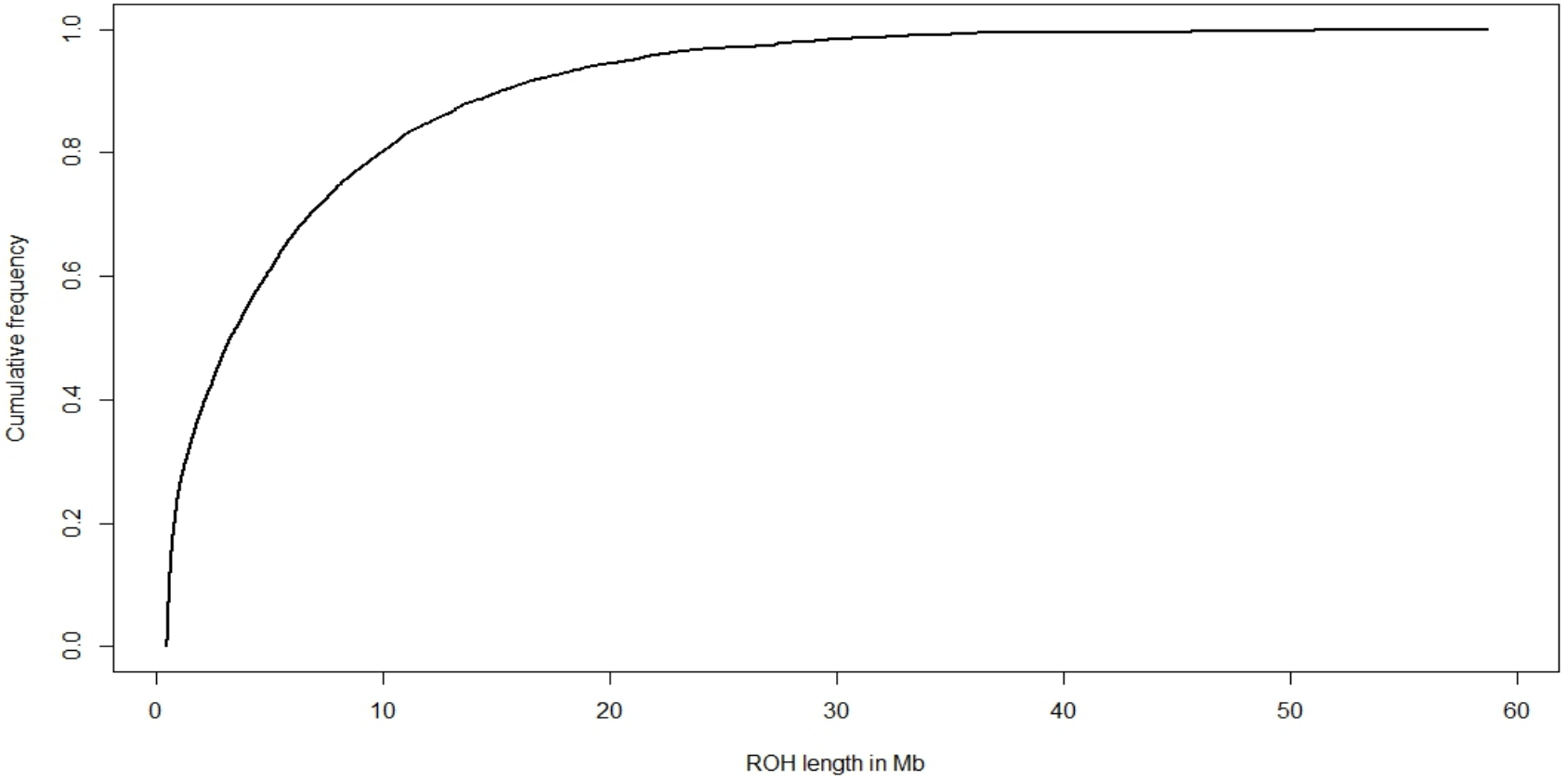


Figure 1



530

531 **Figure 2**