# Efficiency of haplotype-based methods to fine-map QTLs and embryonic lethal variants affecting fertility: illustration with a deletion segregating in Nordic Red cattle

*Naveen Kumar Kadri*[a], *Goutam Sahana*[b], *Bernt Guldbrandtsen*[b], *Mogens Sandø Lund*[b], *Tom Druet*[a,*]

[a]Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), 1 Avenue de l'Hôpital, 4000-Liège, Belgium

[b]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetic, Aarhus University, Blichers Alle20, Postbox50, DK-8830 Tjele, Denmark.

nk.kadri@ulg.ac.be

goutam.sahana@agrsci.dk

bernt.guldbrandsten@agrsci.dk

mogens.lund@agrsci.dk

tom.druet@ulg.ac.be

*Corresponding author: Tel: 003243669172 – Fax: 003243664198

E-mail address: tom.druet@ulg.ac.be

**Abstract**

Despite its importance, fertility has been declining in many cattle populations. In dairy cattle, this decline is often attributed to the negative correlation between fertility and productions traits. Recent studies showed that embryonic lethal variants might also account for a non-negligible fraction of the fertility decline. Therefore identification of such embryonic lethal variants is essential to improve fertility. We herein illustrate, with an example of a large recessive lethal deletion recently identified in Nordic Red cattle, that haplotype-based method are particularly efficient to identify such embryonic lethal variants. We first show that haplotypes can be used in traditional QTL mapping approaches and that they present very high linkage disequilibrium with underlying variants. Haplotypes can also be used in scan for lack of homozygosity. Indeed, if a haplotype is associated to a recessive lethal variant, significantly fewer living individuals will be homozygote for that haplotype than expected. For both approaches, haplotype-based methods were particularly efficient. The lack of homozygosity approach achieved higher significance than the QTL approach. Only frequent variants can be detected with both approaches unless huge genotyped cohorts are available. An alternative approach would rely on identifying potential harmful variants in next-generation sequencing data followed by the genotyping of a larger population for these variants.

**Keywords:** recessive variants, Ancestral Haplotypes, lack of homozygosity, association studies

## Introduction

In recent years, evidence for segregation of recessive embryonic lethal (**EL**) variants (homozygous embryos die) in cattle populations has been reported in several studies (VanRaden et al., 2011; Charlier et al., 2012; Fritz et al., 2013; Sahana et al., 2013; Sonstegard et al., 2013; Kadri et al., 2014; McClure et al., 2014). Such variants probably have a significant contribution to the fertility decline observed in cattle since they reduce the fertility of carriers. Due to the high selection intensity and reduced effective population size in cattle, EL variants can reach higher frequencies than in natural populations where they are under purifying selection.

In the past, such variants were difficult to detect due to the absence in most cases of clinical observations (e.g., calves with malformation). The first identified recessive lethal variants, Complex Vertebral Malformation (CVM) (Agerholm et al., 2001) and Brachyspina (Charlier et al., 2012), correspond to EL associated at least in a few cases with late abortion (and hence observable) although in most cases the embryo dies before birth and the genetic defect remains unobserved. In recent studies (VanRaden et al., 2011; Fritz et al., 2013; Sahana et al., 2013), a new strategy has been used to identify EL variants. It relies on the observation of frequent haplotypes which are never homozygous in live animals or for which the number of homozygous animals is significantly lower than expected (depletion in homozygotes). Such haplotypes potentially carry harmful recessive variants causing death of embryos or animals (including both classical observable genetic defects and EL). This strategy can be implemented when large cohorts of animals are genotyped at high marker density such as those available as a result of the implementation of genomic selection in many dairy cattle populations. Indeed, large numbers of animals are required to find significant depletion in homozygotes for variants segregating at low or moderate frequency. Application of this new strategy in association with fine-mapping tools (based on whole-genome or exome sequencing) has already resulted in the identification of several such EL variants (Sonstegard et al., 2013; Fritz et al., 2013; McClure et al., 2014; Daetwyler et al., 2014).

Recently we identified an embryonic lethal deletion in Nordic Red Cattle with a massive effect on fertility (Kadri et al., 2014). The variant was first identified (Schulman et al., 2011) and fine-mapped as a QTL rather than with

a lack of homozygosity approach. Noteworthy, despite a significant depletion in homozygotes in the Red Danish cattle, it was not significant as a QTL in that population. The fine-mapping was realized with a haplotyped-based method relying on Ancestral Haplotypes (Scheet and Stephens, 2006; Su et al., 2008; Druet and Georges, 2010) although in most cases single-point association studies (SNP-by-SNP) are the preferred method for QTL mapping. The Ancestral Haplotypes method automatically clusters similar haplotypes into groups called 'Ancestral Haplotypes' at each marker position. Contrary to most haplotype-based methods, clustered haplotypes do not need to be identical over a fixed length and can present a few differences corresponding to recent mutations not present in all haplotypes from the group or to genotyping errors.

The objectives of the present study are 1) to illustrate the properties of the Ancestral Haplotypes mapping method and compare them to those of single-point association or fixed length haplotypes and 2) to compare application of QTL mapping approach and lack of homozygosity scan to detect EL. To that end, we will use the 660 kb deletion identified in the Nordic Red cattle population as an example.

## Material and method

***Data.*** The data used for the illustration are described in Kadri et al. (2014). A 660 kb deletion located on chromosome 12 (*Bos Taurus Autosome* - **BTA12**) from positions 20,100,649 to 20,763,116 bp (UMD 3.1 map) was identified as the causative variant underlying a fertility QTL segregating in Nordic Red cattle and previously reported in Schulman et al. (2011). Importantly, the deletion was shown to be an EL variant. In total, respectively 2,855, 4,158 and 6,894 Danish Red (**RDCDNK**), Swedish Red (**RDCSWE**) and Finnish Ayrshire (**RDCFIN**) individuals were genotyped with the Illumina Bovine 50K array and 1,338 markers located on BTA12 (UMD3.1 map) were used for further fine-mapping. For QTL mapping, the phenotype was interval from first to last insemination (**IFL**) (carriers of the deletion have an increased IFL because a larger number of inseminations are necessary to produce a live offspring). It was available for 777 RDCDNK, 1,656 RDCSWE and 2,166 RDCFIN individuals.

***Haplotyping method.*** The haplotypes were reconstructed with the PHASEBOOK software package (Druet and Georges, 2010). The genotypes were first phased using pedigree and linkage information with LinkPHASE and then using DAGPHASE (Druet and Georges, 2010) and Beagle (Browning, 2008). Haplotype reconstruction was performed twice, with and without the five SNPs localized in the deletion (these five SNPs are associated with an inflated genotype miscalling rate and cause haplotype reconstruction errors). The haplotypes were then clustered into 40 Ancestral Haplotypes using HiddenPHASE (Druet and Georges, 2010).

We also used fixed length haplotypes (fixed number of markers) approaches to group haplotypes. Haplotypes identical for 25, 50, 75 or 100 successive SNPs were considered identical-by-descent (**IBD**).

***Measuring association of SNPs, fixed length haplotypes or Ancestral Haplotypes with the deletion.*** We confidently genotyped 1,096, 1,221 and 2,139 animals from the three breeds for the deletion based on SNP genotyping signal intensities, homozygosity and Mendelian inconsistencies for SNPs within deletion as described in Kadri et al. (2014). Thanks to these genotypes, we are able to measure the association between Ancestral Haplotypes, fixed length haplotype or SNPs with the deletion.

We used the coefficient of determination of linear models predicting whether individuals were carriers or not of the deletion based either on all Ancestral Haplotypes or on SNP alleles to assess the predictive ability of both variables and how well they captured the variance associated to the deletion. For SNPs, this is equivalent to the squared correlation between the deletion and the SNP genotype and hence to the LD measured as $r^2$. When comparing haplotype-based methods, we first identified the haplotype group associated with the deletion and then estimated the $r^2$ between that haplotype-group and the deletion.

***QTL mapping method.*** The model used to map the QTL is fully described in Kadri et al. (2014). Briefly, it is a variance components based approach (e.g.,George et al., 2000) relying on a mixed model where the phenotype (de-regressed proofs for interval from first to last insemination in cows) is described by a linear combination of a mean, four principal components (to account for population stratification), a random polygenic effect and a random Ancestral Haplotype effect (as in Druet and Georges (2010) for instance). To test for the presence of a

QTL, this full model is compared with a likelihood ratio test (**LRT**) to a reduced model ($H_0$) without the random Ancestral Haplotype effect. The p-values are obtained by assuming that the LRT is distributed as a chi-square distribution with one degree of freedom. For the single–point association, the random Ancestral Haplotype effect is replaced by a fixed additive SNP effect and the significance was estimated with a t-test.

***Lack of homozygosity mapping.*** To identify significant depletion in homozygotes, we applied a screen testing for significant deviations from Hardy-Weinberg (**HW**) proportions using Ancestral Haplotypes as marker alleles (individuals can either carry 0, 1 or 2 copies of a given Ancestral Haplotypes). One test is performed per Ancestral Haplotype (which results in 40 tests per marker position) and we consider only lack of homozygosity contrary to standard Hardy-Weinberg equilibrium (**HWE**) tests. The same test was repeated with haplotype-groups defined based on fixed length windows or with SNP alleles.

## Results

***Haplotype reconstruction with the Ancestral Haplotypes approach.*** To understand properties of different methods, it is important to know that within the deletion, genotypes of carriers are most often miscalled. Indeed in case of a deletion, individuals carrying the deletion are incorrectly called as homozygotes for the marker alleles carried by the homolog chromosome and haplotypes are consequently incorrect.

Among the 40 Ancestral Haplotypes, we observed that one of them (hereafter called A27) was strongly associated with the deletion (Kadri et al., 2014). Ideally, we want that all carriers of the deletion carry one copy of A27 and vice versa. Figure 1 illustrates with a random subset of 15 haplotypes carrying the deletion how the Ancestral Haplotypes method works. Within the deletion, each of the 15 carriers is associated with the Ancestral Haplotype A27 represented in white (other colors correspond to other Ancestral Haplotypes). We can observe that several of the carriers recombined on both side of the deletion (e.g., haplotypes 3, 6 and 7) and share only a small portions of the haplotype associated with the deletion (as little as only 17 SNPs, including the

deletion, for haplotype 3), still they are correctly assigned to Ancestral Haplotype A27 within the deletion. An approach using fixed size windows would miss carriers recombining within the defined window (the smallest common haplotype to the 15 carriers (excluding the five SNPs from the deletion which are not identical among individuals) from our example contains only two markers). As windows get smaller, the risk of grouping haplotypes which are identical-by-state (**IBS**), but not IBD increases. We can also see that for markers in the deletion, carriers might present different alleles caused by the incorrectly called genotypes (in the example, this is the case for the four first markers in the deletion). Despite the presence of these five noisy SNPs, our model still assigns all carriers to Ancestral Haplotype A27 whereas a fixed size windows approach would fail to group haplotypes carrying the deletion for all windows including any of the five SNPs.

*Figure 1*

The Ancestral Haplotypes model is a Hidden Markov Model where each Ancestral Haplotype has its own emission probabilities (the probability to observe a given marker allele in carriers of that Ancestral Haplotype). In our example, the allelic heterogeneity in the deletion results in emission probabilities that depart from 1 for both alleles (ideally, one of the marker alleles should have a probability of 1). The emission probabilities associated to the Ancestral Haplotype A27 are below 0.999 for only three markers (emission probabilities of 0.981, 0.994 and 0.995) among the 100 markers surrounding the deletion (50 SNPs on each side), which demonstrates that the haplotype associated to the deletion is well defined. On the contrary, the five markers in the deletion have intermediate emission probabilities for Ancestral Haplotype A27 (respectively 0.640, 0.818, 0.251, 0.630 and 0.007) and are not helpful to determine if an individual is carrying the deletion or not. They are rather adding noise and making haplotype assignment more difficult.

***Association of Ancestral Haplotypes, fixed length haplotypes or SNPs and the deletion.*** First we compared the association between SNPs or Ancestral Haplotypes and the deletion (Figure 2). With Ancestral Haplotypes, the coefficient of determination of a linear model predicting whether individuals were carrier or not of the deletion was equal to 0.963, maximizing at position 20,261,439 within the deletion whereas with SNPs, the highest $r^2$ value was equal to 0.475 and corresponded to a SNP located at 18,804,912, 1296 kb from the deletion. The second best SNPs had a $r^2$ value of 0.333 and was located at position 22,219,373 Mb, 1456 kb from the deletion. All remaining SNPs had $r^2$ values below 0.3. The haplotype-based approach was clearly superior to SNPs, for both strength of association and closeness to the causative variant.

*Figure 2*

Table 1 compares clustering of haplotypes associated to the deletion with the Ancestral Haplotypes model and with fixed haplotype length approach with 25 SNPs (approximately 1 to 2 Mb) applied on three positions: centered on the deletion and on both sides of it (to avoid the noise created by the markers from the deletion). It reports the association between carriers of the deletion and carriers of the haplotype associated to the deletion. In total, 16/2123 (Ancestral Haplotypes) and 44/2095 (fixed length approach – before the deletion) individuals were incorrectly/correctly associated. After the deletion, most of the incorrect association with the fixed haplotype length approach results from carriers of the deletion that recombined less than 25 SNPs from the deletion (90 carriers). On the other side (before the deletion), only 10 carriers (out of 439) were not identified as carriers based on the haplotypes due to a recombination (but 34 non-carriers were incorrectly associated with the common haplotype). We also observe that the fixed length haplotype approach is inefficient for windows including the deletion (only 95 carriers sharing a common haplotype of 25 SNPs) whereas the Ancestral Haplotypes approach is robust to both issues (recombination and the deletion) and is able to identify all carriers of the deletion.

In this first comparison, the presence of the five markers from the deletion might create phasing problems (since there are many miss-called genotypes) and make it more difficult to accurately assign haplotypes to the correct Ancestral Haplotype. Windows including the deletion will not correctly group haplotypes carrying the deletion based on IBS fixed length haplotypes. Therefore, only windows close to the deletion (but without any of the five SNPs) might indirectly capture the deletion. Since, the deletion is creating phasing problems and since fixed windows methods are sensitive to it, we re-phased the data excluding the five markers from the deletion. This would also mimic a situation where the causative variant is a SNP.

After exclusion of the five SNPs located in the deletion, the most strongly associated position was located at 20,866,586, only 103 kb from the deletion and the coefficient of determination increased to 0.967 (Figure 2). Table 2 compares the association of the deletion with fixed length haplotypes and Ancestral Haplotypes. The association is improved for both methods (with Ancestral Haplotypes and a fixed length of 25 SNPs). For the fixed length approach with 25 SNPs, haplotypes including the deletion fit well the data now and the number or errors drops from 340 to 24 but they are still some carriers (19) that don't have a complete haplotype in the window due to recombinations. Increasing the size of the windows resulted in more missed carriers (increasing from 19 with 25 SNPs to 99, 163 and 213 with 50, 75 and 100 SNPs) and lower association measured as $r^2$.

***Comparsion of QTL-mapping with SNP and Ancestral Haplotypes.*** Figure 3 presents the QTL fine-mapping curves based on the LRTs in the three breeds. The QTL reaches genome-wide significance level only for Swedish Red and Finnish Ayrshire (no p-value < 0.10 in the region in Danish Red). The LR curves are maximized at position 19,67614 in Finnish Ayrshire with a relatively sharp signal (and at position 18,58491 in Swedish Red). In comparison, the use of SNPs produced a less clear signal with highly significant SNPs dispersed over a larger region (the maximum values are located at positions 31,609,919 in RDCSWE and 24,803,254 in RDCFIN). Figure 3 reports the estimated effect of each Ancestral Haplotype for the three breeds (estimated at the maximum LRT value). In Swedish Red and Finnish Ayrshire the same haplotype is driving the signal with an extremely

negative effect on fertility (increased interval from first to last insemination). That Ancestral Haplotype corresponds to Ancestral Haplotype A27 and is segregating in the three breeds at different frequencies (measured exclusively in animals with phenotype): 0.036 (RDCDNK), 0.141 (RDCSWE) and 0.168 (RDCFIN). The estimated effect (respective solutions are 1.45, 3.74 and 9.14 additional days between first and last insemination for carriers of Ancestral Haplotype A27) is much stronger in Finnish Ayrshire than in other breeds.

*Figure 3*

***Lack of homozygosity mapping.*** We applied a screen testing for significant deviations from HW proportions (considering only lack of homozygosity deviations) using Ancestral Haplotypes as marker alleles (individuals can either carry 0, 1 or 2 copies of a given Ancestral Haplotype). The deletion was detected at high level of significance in RDCFIN and RDFSWE (at positions 19,982,250 and 20,180,276). In RDCDNK, the statistical test ($2.3*10^{-5}$) would not be significant after correction for multiple testing (40 HWE tests per marker) but achieves lower p-values than with the QTL mapping approach (p-value > 0.10).

*Figure 4*

We also performed a similar scan but with fixed length haplotypes (size of 25, 50, 75 and 100 markers), the presence of a recessive lethal effect was clearly identified in RDCSWE and RDCFIN close to the deletion (interval from 17,238,290 to 19,900,184 in both breeds). Using 25 SNPs achieved the most significant p-values (therefore results from haplotypes with 50, 75 and 100 SNPs are not plotted) but with lower significance than with Ancestral Haplotypes. With the fixed length haplotypes, the statistical tests dropped for windows

containing the deletion (due to the incorrect genotype calling) and the tests maximized in the regions flanking the deletion whereas with the Ancestral Haplotype method, the test maximized within the deletion in RDCSWE.

Finally, the same approach was performed using SNP alleles; the achieved significances were lower and maximized further from the deletion (at position 19,406,605 in RDCSWE and 18,804,912 in RDCFIN). The signal was also less sharp than with haplotype-based methods.

**Discussion**

In the present study, haplotypes were in much stronger linkage disequilibrium with the deletion and the association maximized much closer to the causative variant than single-point association studies. They are therefore the preferred method to apply the lack of homozygosity mapping approach as did VanRaden et al. (2011), Fritz et al. (2013) or Sahana et al. (2013). We herein also illustrate that haplotypes can bring additional information compared to SNPs. In our example, the haplotypes allowed us to clearly identify one haplotype with an extremely negative effect on fertility and present in several breeds. The same haplotype presented a strong depletion in homozygotes; this indicated that the variant had a recessive effect (in the offspring). In Kadri et al. (2014), it was also observed that for five consecutive SNPs, individuals carrying the haplotype had a reduced genotyping signal intensity combined with increased homozygosity and Mendelian inconsistencies (genotype incompatibilities in parent-offspring pairs), indicating that a deletion was associated with the haplotype.

In addition, haplotype-based methods offer other benefits. First, haplotype tests can capture allelic heterogeneity (multiple alleles at the same locus) or allelic series, which appears to be common in livestock. For instance, multiple functional mutations have been identified in the myostatin gene (Grobet et al., 1998) or in *MRC2* (Fasquelle et al., 2009; Sartelet et al., 2012), both affecting muscular development in cattle. Furthermore, haplotype models can handle multiple (interacting) mutations at different tightly linked sites. For

instance, comparison in rat of single-point association studies and haplotype-based methods revealed that, although all variants are accurately imputed, haplotypes capture more variation than SNPs (and are more significant), suggesting that many QTL are the result of the combination of multiple linked alleles (Baud et al., 2013).

Associated with imputation techniques (generally relying on haplotypes) single-point association studies are much more efficient, particularly when the causative variants are present in the reference panel used for imputation; a situation that is becoming more common with whole-genome sequencing data (e.g. Daetwyler et al., 2014). In that case, LD can be high (ideally perfect) and allelic heterogeneity is captured. However, first studies on imputation from sequence data suggest that the achieved accuracy is not optimal for all variants and that large cohorts of individuals need to be sequenced, particularly for low frequency variants (e.g. Brøndum, 2013; Daetwyler et al., 2014; Druet et al., 2014). Variants such as the deletion presented in this study might not be systematically imputed (imputation often focuses on SNPs or small indels) or might be more difficult to impute. Finally, haplotypes might also capture rare alleles not present in the reference panel used for imputation.

The method based on Ancestral Haplotypes proved particularly efficient in our illustration. We showed earlier that the method groups haplotypes with high IBD probabilities (as defined by Meuwissen and Goddard (2001)) or with short time of coalescence (Druet and Georges, 2010). In Zhang et al. (2012) or Druet and Farnir (2013), we presented more examples of the high LD between Ancestral Haplotypes and variants (including CNVs) compared to single SNPs. As illustrated, the method uses optimal length of each haplotype to cluster them and is able to group IBD haplotypes even in the presence of genotyping errors, noise or recent mutations (for instance, it was efficient around the deletion although genotypes from carriers were miscalled). As a result, they present high linkage disequilibrium with underlying variants and can even capture CNV (e.g., Durkin et al. (2012); Dupuis et al. (2011)). Ideally, CNV should be modeled more properly with, for instance, the possibility

to emit null or multiple alleles (in case of deletions or duplications). Su et al. (2010) proposed an extension of the HMM to model more appropriately CNV. Small assembly errors might generate similar noise, when a piece of the X chromosome is mapped on an autosome. The Ancestral Haplotype model is not the unique haplotype-based method using variable length haplotypes. For instance, the identity-by-descent approach proposed by Meuwissen and Goddard (2001) or the localized haplotype clustering method from Browning and Browning (Browning and Browning, 2007) do not require haplotypes to be identical over a fixed length.

In our example, the QTL approach was not able to detect the variant in the Red Danish population although we estimated that 13% of the individuals were carriers of the mutation whereas the lack of homozygosity approach achieved more significant statistics. Several parameters explain why the QTL is not detected in that population. First, the Danish data included fewer individuals, only 777 genotyped animals with records compared to 1,656 and 2,166 in Swedish Red and Finnish Ayrshire, respectively. In addition to the number of records, the frequency of a variant also influences the power of QTL detection and the deletion was much less frequent in the Danish population (13 % carriers) compared to the two other breeds (23 and 32% carriers). In the case of EL, the QTL effect is also function of the frequency of carriers in the population, which has therefore a double impact on the power of QTL detection. Indeed the expected effect on Non-Return Rate (NRR) for a carrier of the deletion is proportional to the probability of a carrier producing a homozygote embryo which in turn is proportional to the current allele frequency. For instance, the effect associated to Ancestral Haplotype A27 was a function of the deletion frequency, stronger in Finnish Ayrshire.

In the three breeds, the lack of homozygosity approach was much more significant. The power of the lack of homozygosity approach is a function of the squared difference between observed and expected homozygotes divided by the expected homozygotes (in addition of the deviations for the two other categories). When no homozygotes are observed, the main component of the statistical test is equal to the size of the sample

multiplied by the expected frequency of homozygotes. In the QTL mapping approach, the EL variant is modeled as an additive effect in the parents and this only poorly captures the variance associated with the deletion since it is a recessive effect in the embryo. With the lack of homozygosity mapping, the expected number of homozygotes individuals is used (which more correctly models the data), but a model based on the actual genotype of the embryo (for each insemination resulting in a fertilized egg) would be still more effective although very difficult to obtain. In case of non-random mating, a model based on the genotypes of the parents (only carrier by carrier matings are at risk) would offer a better estimation of the expected number of homozygotes than methods relying simply on frequencies.

*Figure 5*

The significance of the lack of homozygosity approach as a function of the frequency of the variant and the size of genotyped samples is presented in Figure 5 (it is assumed that the $r^2$ between the haplotype and the variant is one). With samples of 1,000 individuals, the significance would be close to $10^{-3}$ and $10^{-7}$ with variant frequencies of 10 and 15%. These are extremely high frequencies (corresponding to 20 and 30% carriers) considering the highly deleterious effect of recessive lethals (the deletion presented in our study segregates at such high frequencies because it is associated to increased milk production). With populations of 10,000 genotyped individuals, the power is much larger but still low when the allele frequency is below 5% (10% carriers). In case of lower LD between the recessive variant and the haplotype, the power would be even lower. Large designs will be required to detect all recessive lethals segregating at low frequency and the genotype-driven screen proposed by Charlier et al. (2012) might then be more efficient. It relies on next-generation sequencing data to screen for coding variants predicted to be disruptive and genotype a larger population for these variants.

## Conclusions

Embryonic lethal variants account probably for a non-negligible fraction of the fertility decline in cattle. We herein illustrated that haplotypes are particularly efficient to identify such variants either by traditional QTL mapping approach or by scan for lack of homozygosity (for a given haplotype). This second approach achieved higher-significance that the first one. However, unless extremely large cohorts of genotyped individuals are available, only frequent variants can be identified. An alternative approach would rely on identifying coding variants predicted to be lethal in sequenced individuals and then to genotype a larger population for these variants.

## Acknowledgements

This work was performed with data from the project 'Genomic Selection – from function to efficient utilization in cattle breeding' (grant no. 3412-08-02253), funded by the Danish Directorate for Food, Fisheries and Agri Business, VikingGenetics, Nordic Cattle Genetic Evaluation, and Aarhus University. The authors thank Carole Charlier for her comments and suggestions on this work. Tom Druet is Research Associate from the Belgian Fond National pour la Recherche Scientifique (FNRS).

## References

Agerholm, J.S., Bendixen, C., Andersen, O., Arnbjerg, J., 2001. Complex vertebral malformation in holstein calves. Journal of veterinary diagnostic investigation : official publication of the American Association of Veterinary Laboratory Diagnosticians, Inc 13, 283-289.

342 Baud, A., Hermsen, R., Guryev, V., Stridh, P., Graham, D., McBride, M.W., Foroud, T., Calderari, S., Diez, M.,
343 Ockinger, J., Beyeen, A.D., Gillett, A., Abdelmagid, N., Guerreiro-Cacais, A.O., Jagodic, M., Tuncel, J., Norin, U.,
344 Beattie, E., Huynh, N., Miller, W.H., Koller, D.L., Alam, I., Falak, S., Osborne-Pellegrin, M., Martinez-Membrives,
345 E., Canete, T., Blazquez, G., Vicens-Costa, E., Mont-Cardona, C., Diaz-Moran, S., Tobena, A., Hummel, O.,
346 Zelenika, D., Saar, K., Patone, G., Bauerfeind, A., Bihoreau, M.T., Heinig, M., Lee, Y.A., Rintisch, C., Schulz, H.,
347 Wheeler, D.A., Worley, K.C., Muzny, D.M., Gibbs, R.A., Lathrop, M., Lansu, N., Toonen, P., Ruzius, F.P., de Bruijn,
348 E., Hauser, H., Adams, D.J., Keane, T., Atanur, S.S., Aitman, T.J., Flicek, P., Malinauskas, T., Jones, E.Y., Ekman,
349 D., Lopez-Aumatell, R., Dominiczak, A.F., Johannesson, M., Holmdahl, R., Olsson, T., Gauguier, D., Hubner, N.,
350 Fernandez-Teruel, A., Cuppen, E., Mott, R., Flint, J., 2013. Combined sequence-based and genetic mapping
351 analysis of complex traits in outbred rats. Nat Genet 45, 767-775.
352 Brøndum, R.F., 2013. Genomic predictions using combined populations and SNP marker panels. PhD thesis,
353 Aarhus University, Faculty of Science and Technology, Aarhus, Denmark.
354 Browning, B.L., Browning, S.R., 2007. Efficient multilocus association testing for whole genome association
355 studies using localized haplotype clustering. Genet Epidemiol 31, 365-375.
356 Browning, S.R., 2008. Missing data imputation and haplotype phase inference for genome-wide association
357 studies. Hum Genet 124, 439-450.
358 Charlier, C., Agerholm, J.S., Coppieters, W., Karlskov-Mortensen, P., Li, W., de Jong, G., Fasquelle, C., Karim, L.,
359 Cirera, S., Cambisano, N., Ahariz, N., Mullaart, E., Georges, M., Fredholm, M., 2012. A deletion in the bovine
360 FANCI gene compromises fertility by causing fetal death and brachyspina. PLoS One 7, e43085.
361 Daetwyler H.D., Capitan A., Pausch H., Stothard P., Van Binsbergen R., Brøndum R.F., Liao X., Djari A., Rodriguez
362 S., Grohs C., Jung S., Esquerré D., Bouchez O., Gollnick N., Rossignol M.N., Klopp C., Rocha D., Fritz S., Eggen A.,
363 Bowman P., Coote D., Chamberlain A., Vantassell C.P., Huggsle I., Goddard M.E., Guldbrandtsen B., Lund M.S.,
364 Veerkamp R., Boichard D., Fries R., Hayes B.J. 2014. The 1000 bull genomes project. Nature Genetics (in press).
365 Druet, T., Farnir, F., 2013. Use of ancestral haplotypes in genome-wide association studies. Methods Mol Biol
366 1019, 347-380.

Druet, T., Georges, M., 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184, 789-798.

Druet, T., Macleod, I.M., Hayes, B.J., 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb) 112, 39-47.

Dupuis, M.C., Zhang, Z., Druet, T., Denoix, J.M., Charlier, C., Lekeux, P., Georges, M., 2011. Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse. Mamm Genome.

Durkin, K., Coppieters, W., Drogemuller, C., Ahariz, N., Cambisano, N., Druet, T., Fasquelle, C., Haile, A., Horin, P., Huang, L., Kamatani, Y., Karim, L., Lathrop, M., Moser, S., Oldenbroek, K., Rieder, S., Sartelet, A., Solkner, J., Stalhammar, H., Zelenika, D., Zhang, Z., Leeb, T., Georges, M., Charlier, C., 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature 482, 81-84.

Fasquelle, C., Sartelet, A., Li, W., Dive, M., Tamma, N., Michaux, C., Druet, T., Huijbers, I.J., Isacke, C.M., Coppieters, W., Georges, M., Charlier, C., 2009. Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet 5, e1000666.

Fritz, S., Capitan, A., Djari, A., Rodriguez, S.C., Barbat, A., Baur, A., Grohs, C., Weiss, B., Boussaha, M., Esquerre, D., Klopp, C., Rocha, D., Boichard, D., 2013. Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. PLoS One 8, e65550.

George, A.W., Visscher, P.M., Haley, C.S., 2000. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics 156, 2081-2092.

Grobet, L., Poncelet, D., Royo, L.J., Brouwers, B., Pirottin, D., Michaux, C., Menissier, F., Zanotti, M., Dunner, S., Georges, M., 1998. Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. Mamm Genome 9, 210-213.

Kadri, N.K., Sahana, G., Charlier, C., Iso-Touru, T., Guldbrandtsen, B., Karim, L., Nielsen, U.S., Panitz, F., Aamand, G.P., Schulman, N., Georges, M., Vilkki, J., Lund, M.S., Druet, T., 2014. A 660-kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in nordic red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet 10, e1004049.

McClure, M.C., Bickhart, D., Null, D., Vanraden, P., Xu, L., Wiggans, G., Liu, G., Schroeder, S., Glasscock, J., Armstrong, J., Cole, J.B., Van Tassell, C.P., Sonstegard, T.S., 2014. Bovine Exome Sequence Analysis and Targeted SNP Genotyping of Recessive Fertility Defects BH1, HH2, and HH3 Reveal a Putative Causative Mutation in SMC2 for HH3. PLoS One 9, e92769.

Meuwissen, T.H., Goddard, M.E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33, 605-634.

Sahana, G., Nielsen, U.S., Aamand, G.P., Lund, M.S., Guldbrandtsen, B., 2013. Novel harmful recessive haplotypes identified for fertility traits in nordic holstein cattle. PLoS One 8, e82909.

Sartelet, A., Klingbeil, P., Franklin, C.K., Fasquelle, C., Geron, S., Isacke, C.M., Georges, M., Charlier, C., 2012. Allelic heterogeneity of Crooked Tail Syndrome: result of balancing selection? Anim Genet 43, 604-607.

Scheet, P., Stephens, M., 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78, 629-644.

Schulman, N.F., Sahana, G., Iso-Touru, T., McKay, S.D., Schnabel, R.D., Lund, M.S., Taylor, J.F., Virta, J., Vilkki, J.H., 2011. Mapping of fertility traits in Finnish Ayrshire by genome-wide association analysis. Anim Genet 42, 263-269.

Sonstegard, T.S., Cole, J.B., VanRaden, P.M., Van Tassell, C.P., Null, D.J., Schroeder, S.G., Bickhart, D., McClure, M.C., 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. PLoS One 8, e54872.

Su, S.Y., Asher, J.E., Jarvelin, M.R., Froguel, P., Blakemore, A.I., Balding, D.J., Coin, L.J., 2010. Inferring combined CNV/SNP haplotypes from genotype data. Bioinformatics 26, 1437-1445.

Su, S.Y., Balding, D.J., Coin, L.J., 2008. Disease association tests by inferring ancestral haplotypes using a hidden markov model. Bioinformatics 24, 972-978.

VanRaden, P.M., Olson, K.M., Null, D.J., Hutchison, J.L., 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J Dairy Sci 94, 6153-6161.

Zhang, Z., Guillaume, F., Sartelet, A., Charlier, C., Georges, M., Farnir, F., Druet, T., 2012. Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Bioinformatics 28, 2467-2473.

Figure 1. Illustration of Ancestral Haplotypes clustering for a set of 15 haplotypes (one per line) carrying the deletion. The represented region contains 50 markers and is centered on the deletion (the red vertical lines represent its borders). Colors correspond to the Ancestral Haplotype to which each haplotype was assigned at that position (white for Ancestral Haplotype A27). Numbers (1 or 2) correspond to the marker allele carried by the haplotype at that position. We can observe that similar haplotypes are grouped in the same Ancestral Haplotype and that a same haplotype can switch from one Ancestral Haplotype to another at any position along the chromosome.

Figure 2. Association with the deletion, measured as the coefficient of determination of a linear model using SNPs (gray dots), Ancestral Haplotypes (black line) or Ancestral Haplotypes after removal of the five SNPs located in the deletion (gray line).

Figure 3. Haplotyped-based QTL fine-mapping and single-point association results on BTA12 with 40 Ancestral Haplotypes for interval from first to last insemination of cows in Finnish Ayrshire (grey line and ●), Swedish Red (black line and +) and Danish Red (black dashed line and •) cattle (left panel). The borders of the deletion are marked by the vertical dashed lines. Effect (in days) and frequency of the 40 Ancestral Haplotypes in Finnish Ayrshire (top), Swedish Red (center) and Danish Red (bottom) cattle (right panel).

Figure 4. Comparison of lack of homozygosity mapping along BTA12 with Ancestral Haplotypes (solid line), with fixed length haplotype windows of 25 markers (dashed line) and with SNPs (gray dots) in Danish Red (top), Swedish Red (center) and Finnish Ayrshire (bottom) cattle. The borders of the deletion are marked by the vertical dashed lines.

443 Figure 5. Significance of deviation from Hardy-Weinberg proportions test as a function of the size of the geno-

444 typed population and the frequency of the recessive variant (allele frequency: 0.01 (-), 0.02 (○), 0.05 (△),

445 0.10(+), 0.15 (■), 0.20 (□)).

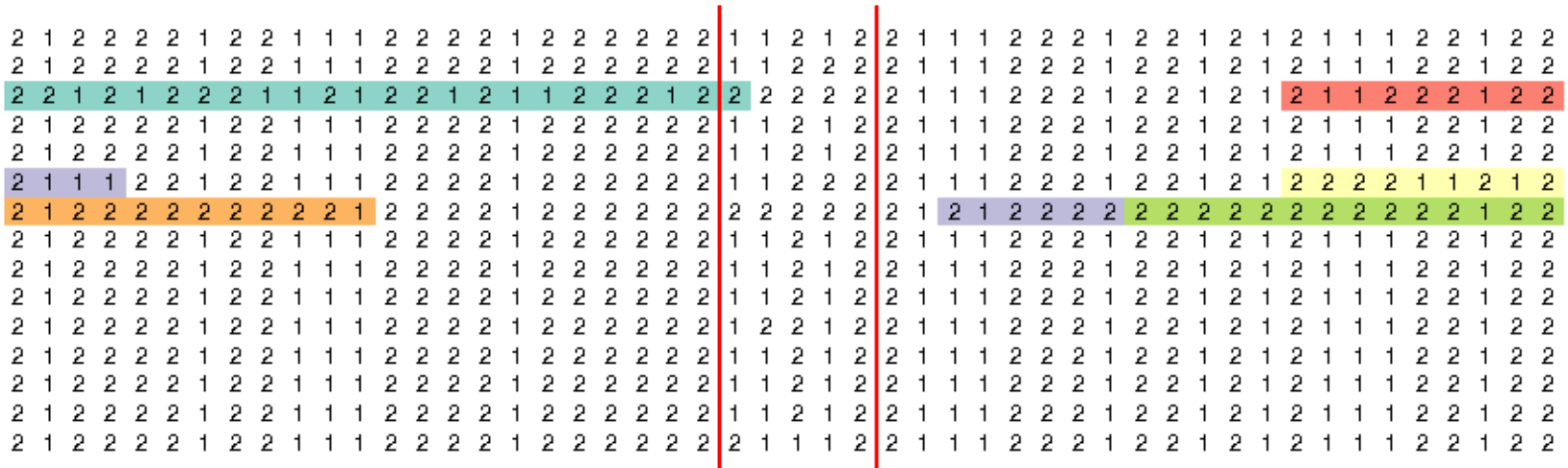
446

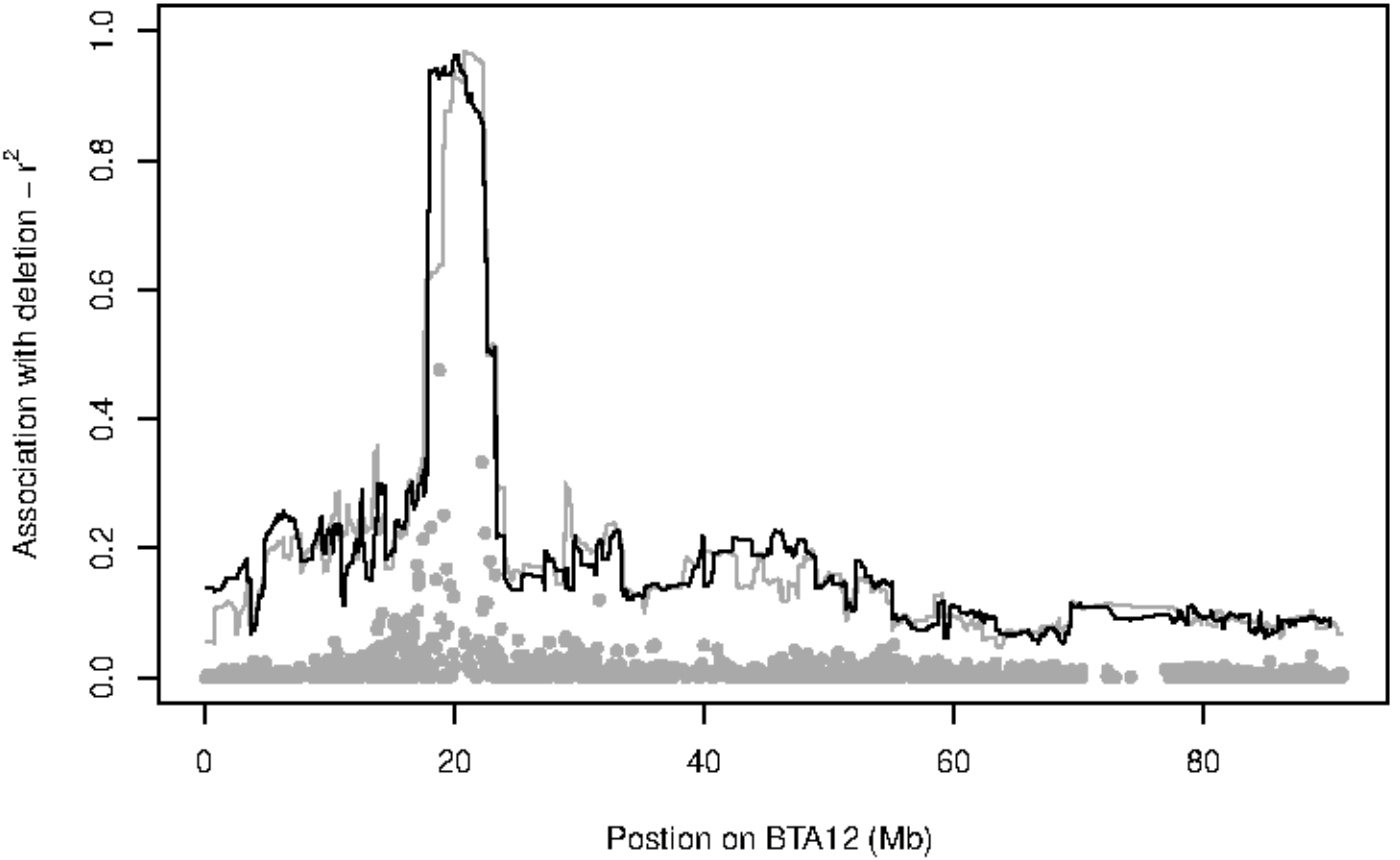
Table 1. Association between carriers of the deletion and carriers of haplotypes associated to the deletion. The comparison was performed in Finnish Ayrshire.

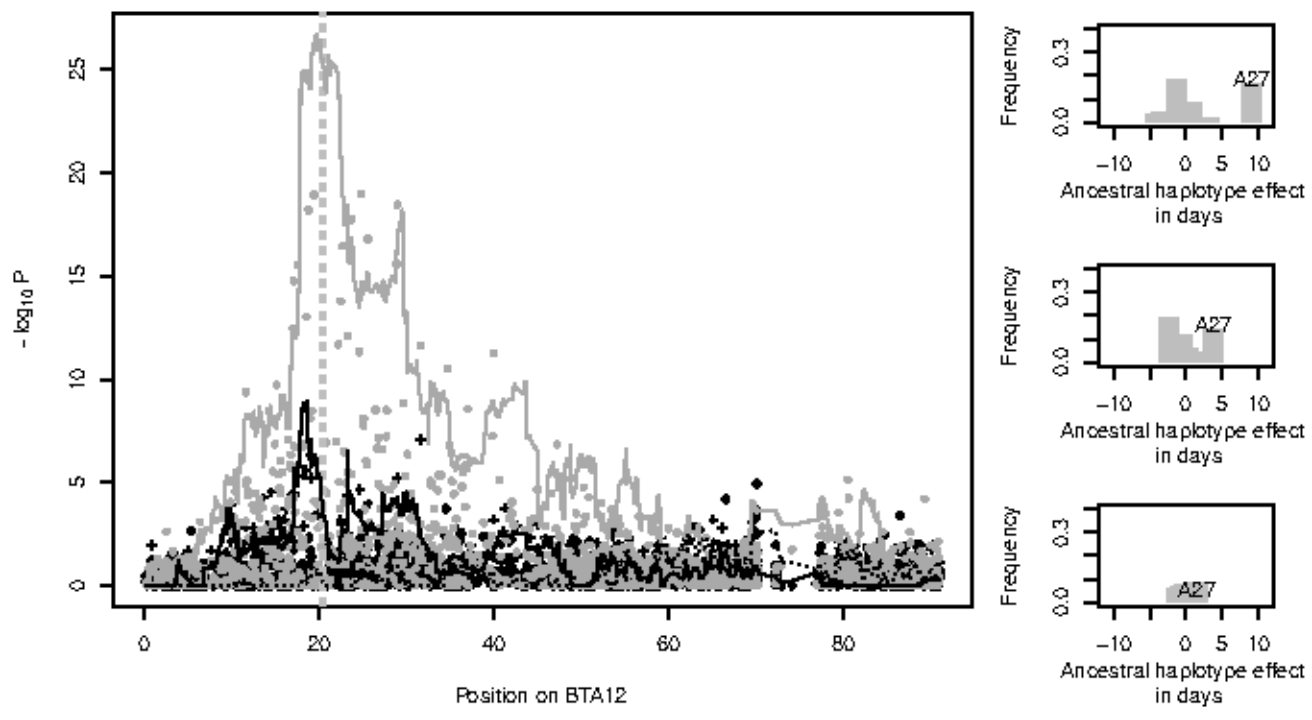
| Method | Position | Carrier of the deletion | | Non-carrier of the deletion | | Accuracy |
|---|---|---|---|---|---|---|
| | | With common haplotype | Without common haplotype | With common haplotype | Without common haplotype | (correlation$^2$) |
| Ancestral Haplotypes | Within the deletion | 439 | 0 | 16 | 1684 | 0.96 |
| Fixed length haplotypes | Centre of deletion | 95 | 344 | 0 | 1700 | 0.18 |
| Fixed length haplotypes | Before deletion | 429 | 10 | 34 | 1666 | 0.88 |
| Fixed length haplotypes | After deletion | 349 | 90 | 14 | 1686 | 0.72 |

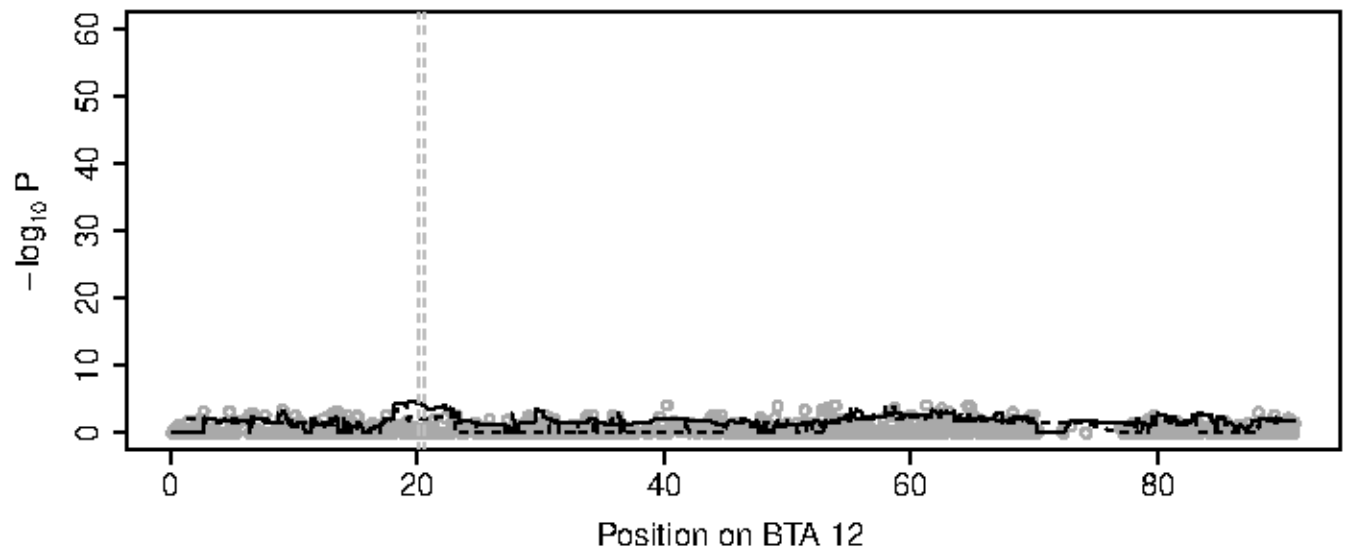
Table 2. Association between carriers of the deletion and carriers of haplotypes associated to the deletion after rephasing the data without the five SNPs located in the deletion. The comparison was performed in Finnish Ayrshire.

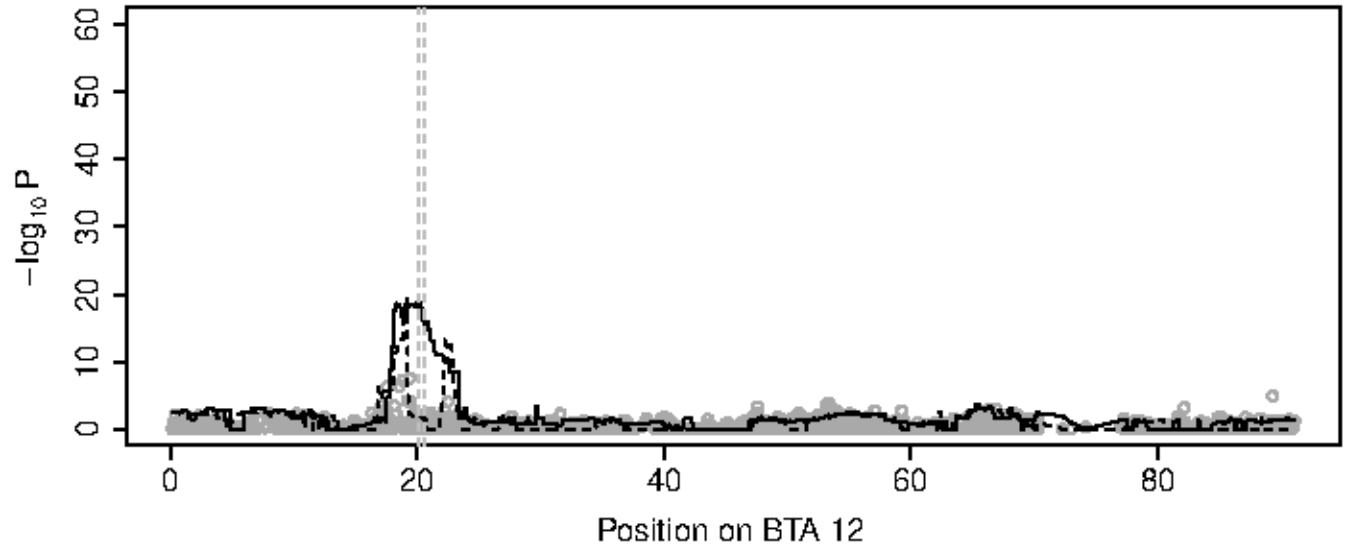
| Method | Number of SNPs | Carrier of the deletion | | Non-carrier of the deletion | | Accuracy (correlation$^2$) |
|---|---|---|---|---|---|---|
| | | With common haplotype | Without common haplotype | With common haplotype | Without common haplotype | |
| Ancestral Haplotypes | / | 428 | 11 | 2 | 1698 | 0.96 |
| Fixed length haplotypes | 25 | 420 | 19 | 5 | 1695 | 0.93 |
| Fixed length haplotypes | 50 | 340 | 99 | 1 | 1699 | 0.73 |
| Fixed length haplotypes | 75 | 276 | 163 | 1 | 1699 | 0.57 |
| Fixed length haplotypes | 100 | 226 | 213 | 1 | 1699 | 0.45 |

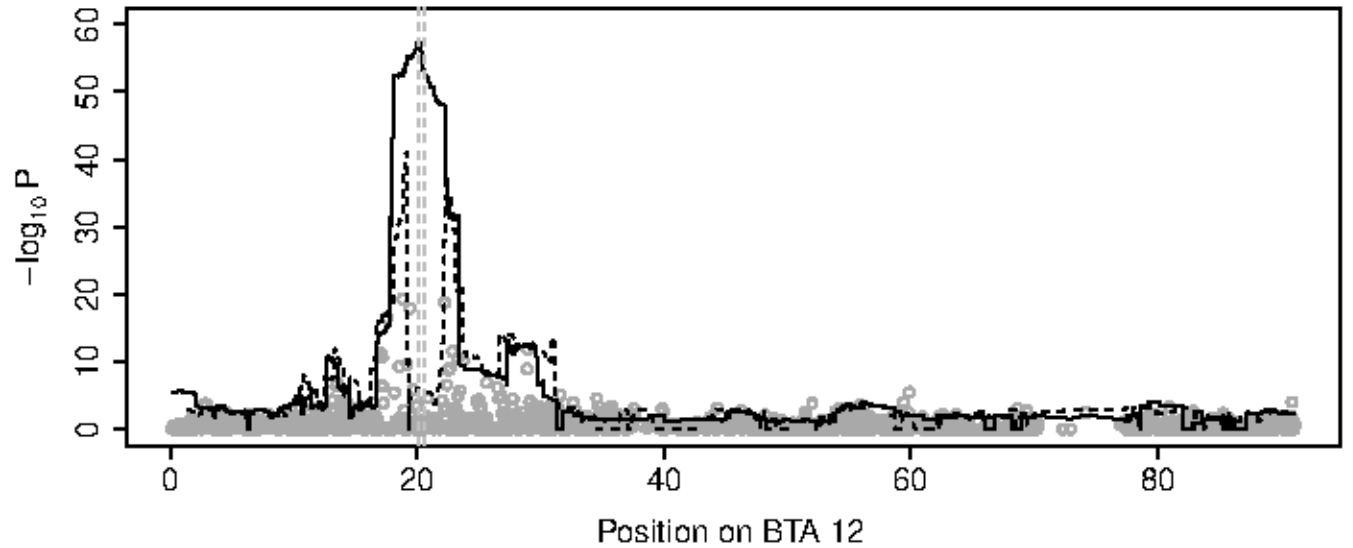


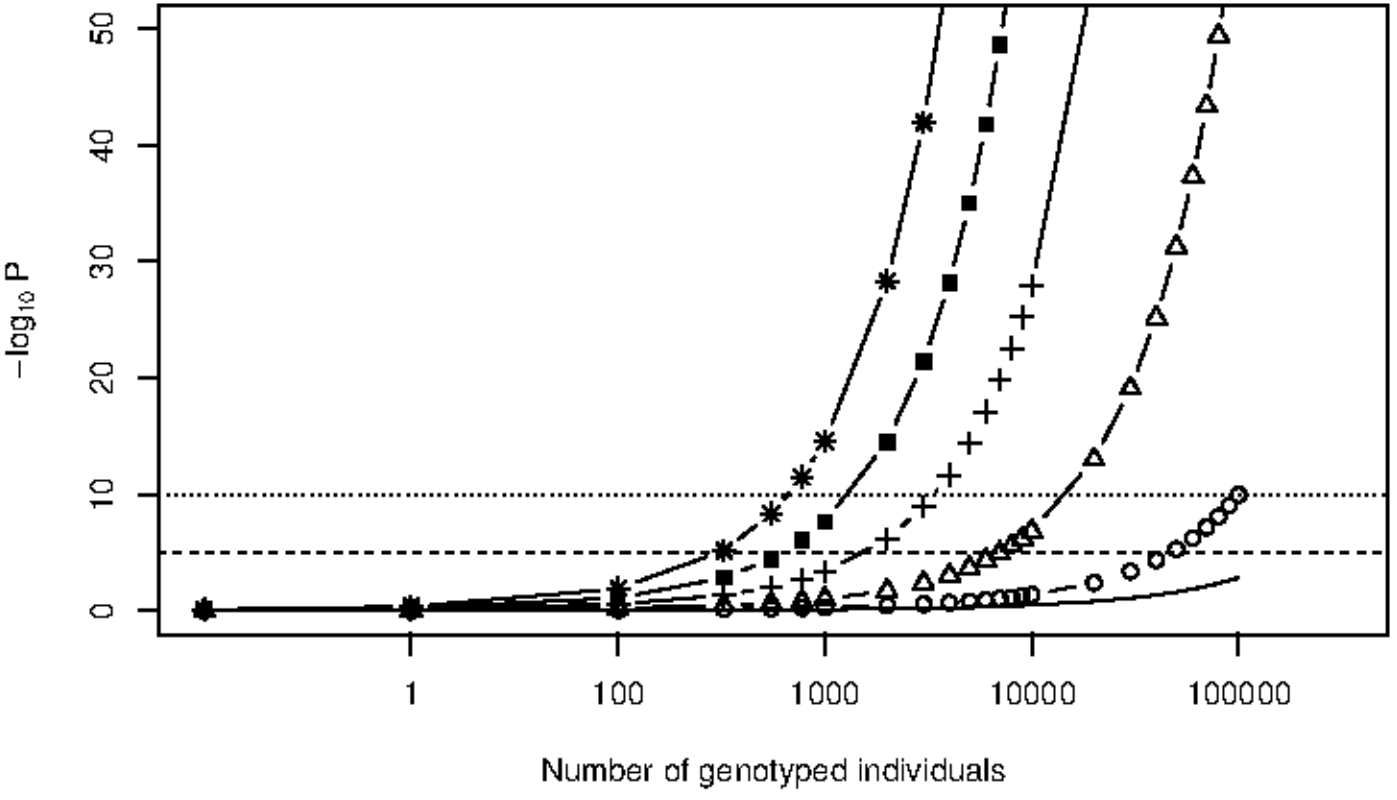

Association with deletion – $r^2$
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Postion on BTA12 (Mb)


- log10 P
0
5
10
15
20
25
0
20
40
60
80
Position on BTA12
Frequency
0.0
0.3
A27
−10
0
5
10
Ancestral haplotype effect in days
Frequency
0.0
0.3
A27
−10
0
5
10
Ancestral haplotype effect in days
Frequency
0.0
0.3
A27
−10
0
5
10
Ancestral haplotype effect in days


−log10 P
0
10
20
30
40
50
60
0
20
40
60
80
Position on BTA 12


−log10 P
0
10
20
30
40
50
60
0
20
40
60
80
Position on BTA 12


−log10 P
0
10
20
30
40
50
60
0
20
40
60
80
Position on BTA 12


−log10 P
0
10
20
30
40
50
1
100
1000
10000
100000
Number of genotyped individuals