Spanish Journal of Agricultural Research
15 (4), e0407, 10 pages (2017)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2017154-11736
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, O.A, M.P. (INIA)

RESEARCH ARTICLE

OPEN ACCESS

A haplotype regression approach for genetic evaluation using sequences from the 1000 bull genomes Project

Kenza Lakhssassi and Oscar González-Recio

Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, O.A., M.P. (INIA). Dept. Mejora Genética Animal. Ctra. de la Coruña, km 7,5 - 28040 Madrid, Spain.

Abstract

Haplotypes from sequencing data may improve the prediction accuracy in genomic evaluations as haplotypes are in stronger linkage disequilibrium with quantitative trait loci than markers from SNP chips. This study focuses first, on the creation of haplotypes in a population sample of 450 Holstein animals, with full-sequence data from the 1000 bull genomes project; and second, on incorporating them into the whole genome prediction model. In total, 38,319,258 SNPs (and indels) from Next Generation Sequencing were included in the analysis. After filtering variants with minor allele frequency (MAF< 0.025) 13,912,326 SNPs were available for the haplotypes extraction with findhap.f90. The number of SNPs in the haploblocks was on average 924 SNP (166,552 bp). Unique haplotypes were around 97% in all chromosomes and were ignored leaving 153,428 haplotypes. Estimated haplotypes had a large contribution to the total variance of genomic estimated breeding values for kilogram of protein, Global Type Index, Somatic Cell Score and Days Open (between 32 and 99.9%). Haploblocks containing haplotypes with large effects were selected by filtering for each trait, haplotypes whose effect was larger/lower than the mean plus/minus 3 times the standard deviation (SD) and 1 SD above the mean of the haplotypes effect distribution. Results showed that filtering by 3 SD would not be enough to capture a large proportion of genetic variance, whereas filtering by 1 SD could be useful but model convergence should be considered. Additionally, sequence haplotypes were able to capture additional genetic variance to the polygenic effect for traits undergoing lower selection intensity like fertility and health traits.

Additional keywords: full sequence; Holstein; findhap.f90; Bayesian model.

Abbreviations used: BTA (Bos Taurus Autosome); DO (Days Open); GEBV (Genomic Estimated Breeding Value); GTI (Global Type Index); GWAS (Genome Wide Association Studies); LD (linkage disequilibrium); MACE (Multiple-trait Across Country Evaluation); MAF (minor allele frequency); PROT (kg of protein); QTL (Quantitative Trait Loci); SCS (Somatic Cell Score); SD (standard deviation); SNP (Single Nucleotide Polymorphism).

Authors' contributions: Conceived and designed: OGR; analysis and interpretation of data; drafting of the manuscript: KL; supervising the work: OGR.

Citation: : Lakhssassi, K.; González-Recio, O. (2017). A haplotype regression approach for genetic evaluation using sequences from the 1000 bull genomes Project. Spanish Journal of Agricultural Research, Volume 15, Issue 4, e0407. https://doi.org/10.5424/sjar/2017154-11736.

Received: 19 May 2017. **Accepted:** 09 Jan 2018.

Funding: The authors received no specific funding for this work.

Competing interests: no competing interests exist.

Correspondence should be addressed to Oscar González-Recio: gonzalez.oscar@inia.es

Introduction

New technological advances such as Single Nucleotide Polymorphism (SNP) discovery using high-throughput SNP genotyping has led to a new strategy of selection called genomic selection (GS) that has revolutionized breeding in some species such as dairy cattle in the last decade (Hayes *et al.*, 2009; Ibañez-Escriche & Gonzalez-Recio, 2011). Genomic predictions are now used routinely in selection of dairy cattle. The achievable genetic gain is proportional to the accuracy of predictions, which depends on the proportion of the genetic variance that can be captured by the marker information. This is function of the linkage disequilibrium (LD) between the SNP and the causative mutations affecting the trait, size of the genotyped population, heritability of the traits, among other factors (Druet *et al.*, 2014). Modern deep sequencing technology offers the possibility to obtain whole genome sequence data, which are expected to include the causal mutations responsible for trait variation (Meuwissen & Goddard, 2010). Consequently, predictions should no longer depend on LD between SNPs and quantitative trait loci (QTL). According to MacLeod *et al.* (2013), inclusion of the causal mutations allows the effect of the QTL on a given trait to be estimated directly,

which should increase the reliability of genomic predictions compared to using SNP genotypes, as well as the persistency of the reliability of predictions across generations. Druet *et al.* (2014) showed that the accuracy of genomic breeding value may improve in the range of 2-30% (depending on the trait), if the variation from rare alleles could be captured from the whole genome sequence data and exploited in genomic predictions. Nonetheless, sequencing many individuals is still too expensive, and imputation to sequence data using SNP genotypes is an attractive and cost-effective approach to obtain a large training set of sequenced individuals. Whole genome imputation accuracy is larger than 0.95, except for variants with minor allele frequency (MAF)>0.10, which can be as low as 0.50 (Van Binsbergen *et al.*, 2014).

The 1000 bull genome project (http://www.1000bullgenomes.com) aims at sequencing a number of key ancestor bulls in the beef and dairy cattle population on an international collaboration between scientists in Europe, USA, and Australia. The National Institute of Investigation and Agricultural and Food Technology (INIA) in Spain joined this consortium since 2015. More than 1,500 animals have been already sequenced in this project, and 31.8 million variants detected (Daetwyler *et al.*, 2014). Sequence data from these animals allows imputing lower density SNP genotypes from other animals to whole sequence, allowing more accurate genome wide association studies (GWAS) and genomic predictions. The amount of information to be analyzed in this situation poses new challenges from a statistical and a computational point of view. The main challenges at dealing with whole genome sequence data for genomic prediction are: the imputation accuracy of low MAF sequence variant, the computational burden and finding proper statistical methods to deal with the high dimensionality problem (Hayes *et al.*, 2014). The main objective of this work was to develop strategies to incorporate sequence information in genetic evaluations using sequences of the 1000 bull genomes project. Firstly, we evaluated the performance of the findhap software to construct haplotypes from sequence data. Secondly, we detected sequence regions that are associated to traits of economic interest and can be incorporated in genomic evaluations in the Spanish dairy cattle. Finally, we evaluated the proportion of genetic variance that can be explained by these regions.

Material and methods

Data

The reference population consisted of 450 Holsteins sires with whole-sequence data from the 1000 Bull Genomes Project (Run 5). In total, 38,319,258 SNPs and indels from Next Generation Sequencing were included. However, a large percentage (50%) of these variants with low MAF are expected to be sequencing errors (Gonzalez-Recio *et al.*, 2015). Variants with MAF< 0.025 were discarded in this study. The number of SNPs remained on each Bos taurus autosomal chromosomes (BTA) is given in Table 1.

Four economically important traits were used in this study: kilograms of protein (PROT), Global Type Index (GTI), Somatic Cell Score (SCS) and Days Open (DO). The phenotypic values were the Multiple-trait Across Country Evaluation (MACE) proofs provided by the Spanish Holstein Association CONAFE. Only animals with sequence and phenotype were kept for further analyses, which left 361 sires with genomic and phenotype information for the subsequent analyses. All available pedigree data were incorporated in the study, in total 435 animals in the pedigree.

Estimation of haplotypes in the population

Haplotypes were obtained from version 3 of findfhap.f90 software (VanRaden *et al.*, 2011). Findhap was performed considering the values recommended by the author, except the error rate parameter and the haplotype length. The error rate parameter is defined as the expected percentage of variants that are incorrectly called at sequencing. At a very large numbers of variants sequenced, the number of sequencing errors is likely to be considerable. Findhap program suggest 0.002 as error rate, but a recent study showed that at MAF<0.01 up to 50% of variants are sequencing errors (Gonzalez-Recio *et al.*, 2015). This creates haplotypes that appear only in one animal (singletons) and thus are not informative. These authors also estimated a sequencing error in variant calling of 1%. Hence, the error rate in this study was set to 0.01.

The haplotype length is defined as the number of SNP contained in the block (haploblock), and is a findhap parameter provided by the user. This is one of the main parameters that need to be determined at implementing the algorithm. A previous study on haplotyping in German Holstein cattle reported a mean block length of 164 kb (Qanbari *et al.*, 2010). A proper definition of the haplotype length will minimize the probability of recombination within the block, and maximize the probability of transmitting the whole block to the progeny. Hence, the number of haplotype blocks and the haplotype length per chromosome were defined as follows:

$$\text{Number of blocks} = \frac{\text{Chromosome length (kb)}}{\text{Block length (kb)}}$$

$$\text{Number of SNP per block} = \frac{\text{Number of SNP remained}}{\text{Number of blocks}}$$

Table 1. Total and filtered SNP (MAF< 0.025) on each *Bos taurus* autosomal chromosome (BTA)

Chromosome number	Total SNP	SNP remained
BTA1	2415624	859904
BTA 2	2062223	692662
BTA 3	1747334	620585
BTA4	1840583	672654
BTA5	1790983	663521
BTA6	1773469	679959
BTA7	1610969	571225
BTA8	1622084	573042
BTA9	1555596	566141
BTA10	1532614	575132
BTA11	1546812	559737
BTA12	1667137	711845
BTA13	1236102	409838
BTA14	1234979	428478
BTA15	1415166	522191
BTA16	1291009	427038
BTA17	1157679	447303
BTA18	964483	361249
BTA19	929690	334276
BTA20	1121685	394257
BTA21	1088553	379736
BTA22	892683	305698
BTA23	1016377	387518
BTA24	994429	346706
BTA25	670204	240877
BTA26	779371	286624
BTA27	698131	286641
BTA28	772863	276837
BTA29	890426	330652

where the average block length was considered 164 kb as proposed by Qanbari *et al.* (2010). Then, haplotype blocks were built separately for each BTA. The options in findhap were set to a minimum length of 800 SNPs, and a maximum length of 100,000 SNPs, with 5 iterations per imputation step. Haplotype alleles with frequency <1% were excluded to eliminate singletons and too low frequency alleles. After this filtering, 153,428 haplotypes were kept for subsequent analyses. Each haplotype was identified by location chromosome and segment as well as the ordered number of the haplotype within the segment. Haplotype data were coded as 0, 1 or 2 according to the number of haplotype alleles that the animal carries. The haplotype data were then merged with the phenotype file for subsequent analyses.

Incorporating sequence haplotypes in the whole sequence prediction model

The following linear equation represents the relationship between the phenotype of interest and the genetic effects (haplotype variants and polygenic effect):

$$y = 1'\mu + Wh + Zg + e$$

where y is a vector of phenotypic observations, μ is a population mean, 1' is a vector of ones, h is the vector of haplotype effects assumed to be distributed as a double exponential (Laplace distribution DE) $h \sim DE(\mu_h,\lambda)$. The λ parameter is a smoothing parameter controlling the shrinkage of the double exponential distribution; λ^2 is distributed a priori as a gamma distribution with a

shape and scale hyperparameters, which were set by a grid search with values ranging from 0.0000001 to 1.

Then, g is the vector of polygenic effects distributed as g ~ N(0,$G\sigma_g^2$), G is the genomic relationship matrix built from Illumina Bovine 50K genotypes. Pairs of individuals sharing the same genotype for a large number of markers will be more similar genomically, and will have higher values in the corresponding off diagonal cells of the matrix, as is the case for pairs of related animals in a pedigree based relationship matrix. The genomic relationship matrix was computed as:

$$G = \frac{ZZ\prime}{2\sum p_i(1-p_i)}$$

where p_i is the frequency of the minor allele at locus *i*, Z = (M - P) is a matrix that results from the subtraction of P from M, being P = 2(p_i − 0.5), and M the matrix of genotypes codified as –1, 0, and 1 for the homozygote, heterozygote, and other homozygote, respectively, following VanRaden (2008), where a more detailed description of this model is provided. The W and Z are the corresponding incidence matrices, and e is the vector of random residual terms of the model e ~ N(0,Dσ_e^2), weighted by the MACE proof accuracy as proposed by De Los Campos *et al.* (2013), being D a diagonal matrix with elements $\{\frac{1-r_i}{r_i}\}$, where r_i is the reliability of individual *i*. Finally, σ_g^2 and σ_e^2 are the additive polygenic and residual variances, respectively.

The Bayesian model was solved for each chromosome separately using Gibbs Sampling, with a chain length of 10,000 and a burn-in period of 1000. It should be noted that the total Genomic Estimated Breeding Value (GEBV) obtained from the prediction models consisted of the sum of the estimated haplotype effects and the polygenic effect estimate as:

$$GEBV = (\sum haplotype\ effects) + polygenic\ effect$$

Haplotype selection

Cuyabano *et al.* (2015) reported that an appropriate selection of a subset of haplotype blocks can result in similar or better predictive ability than using the whole set of haplotype blocks. This was also expected to reduce the dimensionality of the models. Hence, haplotypes were filtered by their estimated effect. Alleles whose estimate, in absolute value, were above 3 (1) SD above the mean were selected for each trait.

$$|\hat{h}| > \mu h + 3\ SD_h$$
$$|\hat{h}| > \mu h + SD_h$$

Then, each analysis was repeated incorporating only haplotypes that exceeded each threshold using the model described above for all chromosomes simultaneously. Genetic variance explained by the selected haplotypes from sequence data was estimated for each trait. The haplotypic genetic variance was estimated as the ratio of variance explained by haplotype over the total GEBV variance.

Results

Haplotypes from sequence data

The length and number of the haplotype segments varied depending on the extent of LD present and on the chromosome length. Table 2 shows summary statistics of the haploblocks found for the 29 BTAs. The total BTA genome length was 2512.06 Mb with the shortest length being 42.90 Mb (BTA25) and the longest one being 158.33 Mb (BTA1). The number of SNPs per haploblock ranged from 799 in BTA13 to 1285 in BTA12, with a mean of 924 SNP (length 166,552 bp). The BTA1 showed the highest number of haplotype blocks (961) as well as haplotypes (9363), while BTA25 showed the smallest number of blocks (261) and haplotypes (2788). Ninety per cent of haplotypes showed only one occurrence (singletons) whereas 97% of haplotypes were present in MAF<1%. These haplotypes were not used in this analysis due to the difficulty of finding statistical effects when the haplotype is present in only a couple of individuals in our sample. Then, low frequency haplotype alleles (<1%) were ignored leaving 153,428 haplotypes for subsequent analysis. It must be pointed out that these rare haploblocks may be still of interest in data sets with larger sample size.

Figure 1 shows the number of genome-wide haplotype blocks against the number of haplotypes remaining after filtering. The distribution of haplotype blocks is proportional to the number of haplotypes. The larger the number of blocks the larger the number of haplotypes. This suggests that the genetic variability within chromosome was proportional to the chromosome length, and we could not detect any chromosome with larger or lower variability than expected, despite that Holstein breed has undergone strong selection intensity during many decades.

Alternative block lengths were also analysed: 100,000 and 2,000 SNP for the maximum and minimum length respectively, as recommended by VanRaden *et al.* (2011). However, it resulted on too long haploblocks, with a large proportion of singletons that were filtered out during the process. In this case, the remained haplotypes (76,512) explained only a small percentage of the variance of the GEBV for each trait.

Table 2. Genome-wide summary of haplotype blocks in the Holstein cattle of this study.

Autosome	Autosome length (bp)	Haplotype number	Blocks number	Blocks length (SNP)	Unique haplotypes (%)	Haplotype freq <1% (%)	Haplotypes remained
BTA1	158334731	397685	961	895	90.26	97.65	9363
BTA2	137060366	335343	830	835	89.69	97.64	7930
BTA3	121430266	302697	735	844	90.27	97.63	7179
BTA4	120825133	303697	736	914	89.78	97.39	7914
BTA5	121190985	314912	738	899	90.68	97.71	7224
BTA6	119458581	300751	724	939	90.15	97.59	7237
BTA7	112638649	285068	685	834	90.84	97.78	6328
BTA8	113383722	285048	687	834	90.31	97.62	6772
BTA9	105708161	272927	644	879	90.40	97.64	6446
BTA10	104304932	259351	633	909	89.25	97.58	6277
BTA11	107310498	272252	651	860	90.11	97.50	6807
BTA12	91163122	252401	554	1285	86.26	97.78	5597
BTA13	84240314	212863	513	799	90.48	97.79	4707
BTA14	84648338	206927	514	834	89.48	97.55	5060
BTA15	85295694	218218	518	1008	90.13	97.45	5565
BTA16	81724537	205580	497	859	90.40	97.74	4655
BTA17	75158596	197683	457	979	90.54	97.72	4507
BTA18	66003508	175914	402	899	90.66	97.50	4399
BTA19	64057258	166221	389	859	90.56	97.65	3910
BTA20	72041471	180605	439	898	89.87	97.69	4164
BTA21	71599084	183741	434	875	90.66	97.61	4398
BTA22	61435160	152167	373	820	89.49	97.45	3879
BTA23	52529233	137483	319	1215	89.30	97.45	3507
BTA24	62714571	155380	381	910	89.52	97.46	3953
BTA25	42904110	108716	261	923	89.26	97.44	2788
BTA26	51680365	128747	314	913	89.17	97.53	3182
BTA27	45407501	116888	276	1039	89.02	97.27	3195
BTA28	46312540	118038	282	982	89.37	97.33	3157
BTA29	51505224	134472	312	1060	90.57	97.53	3328

BTA : *Bos taurus* autosomal chromosome

Contribution of haplotype effects on the phenotypes

Graphical representation of the haplotypes effect estimates are shown for each trait in Figure 2. Each BTA is represented in a different color. The genome wide threshold of 3 SD is shown as a horizontal line. We detected some regions on the genome that explained a relevant effect for the studied traits. As summary, 1264 haplotypes exceeded the genome wide threshold for PROT, 1909 for GTI, 851 for SCS and 1450 for DO. The chromosome 1 was the one with the greatest number of haplotypes which effect exceeded the 3 SD threshold, for all the traits (132 for PROT, 199 for GTI, 94 for SCS and 140 for DO).

Figure 3 shows the distribution of haplotype allelic frequencies that exceeded the 3 SD threshold for each trait. Most of the haplotypes for PROT, GTI and DO had low-intermediate frequencies while haplotypes found for SCS seemed to be at even lower frequencies. These haplotypes may provide additional information to SNP genotypes for less common variants, because SNP beadchips were designed to genotype intermediate-high MAF.

Genetic variance explained by sequence data

Table 3 shows the proportion of GEBV variance explained by the sequence haplotypes. Haplotypes

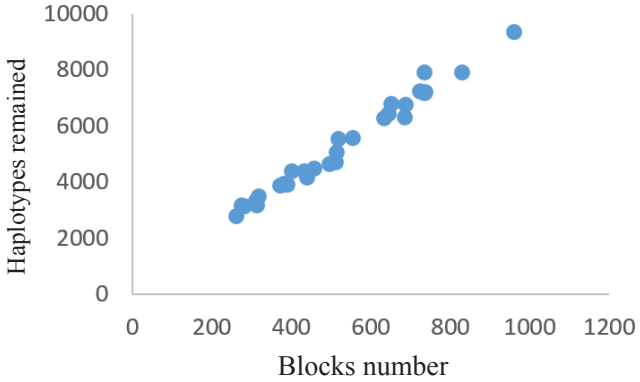

Figure 1. Scatter plot for the number of blocks against the number of haplotype remained after filtering for each chromosome

explained a large proportion of the total GEBV variance (32-99%). Smaller contribution of haplotypes to genetic variance resulted for the production trait (PROT), whereas they captured almost all the variance for SCS, probably as an artifact caused by data structure, the large *p* small *n* problem, larger proportion of missing data or the lower heritability of the trait.

Two subsets of haploblocks were selected to reduce the number of unknowns needed to perform genomic prediction. The haplotypes were selected according to the magnitude of the effect estimate. The first subset contained haplotypes with effect estimates (in absolute value) above 3 SD above the mean. This led to 1264 haplotypes for PROT, 1909 for GTI, 851 for SCS and 1450 for DO. The second subset contained haplotypes with effect estimates (in absolute value) larger than 1 SD above the mean, leading to a total of 44,319 haplotypes for PROT, 39,975 for GTI, 46,132 for SCS and 42,878 for DO. Then, each subset was subjected to a new analysis with Bayesian LASSO.

As expected, the larger the number of haplotypes the larger the proportion of genetic variance that was captured by sequence haplotypes. The subset of haplotypes with larger effect (>3 SD) contained only between 1 and 5 % of haplotypes in the larger subset (>1 SD). Despite this low number of covariates, they explained up to half the variance captured by haplotypes over 1 SD (10% for kg PROT, 22% for DO and 46% for SCS). We did not obtain convergence for GTI with the threshold of 1 SD in the case of PROT, selected haplotypes with larger effect estimate (>3 SD) might be pointing to few genomic regions strongly associated to the trait, and with a large number of haplotypes each, but not representative of the whole genetic architecture (failing to identify/select other regions).

A G-BLUP model without including haplotype effects was implemented as benchmark, to evaluate the contribution of sequence haplotypes at capturing additional genetic variance. Table 4 shows that the polygenic variance decreased for all traits when sequence haplotypes were included in the model. This reduction ranged between 41 % (PROT) and 83 % (DO), and got accentuated with larger number of haplotype effects included in the model. The reduction was larger for fertility and mastitis related traits, and lower for the production and conformation traits. Residual variance also decreased when haplotypes were included in the

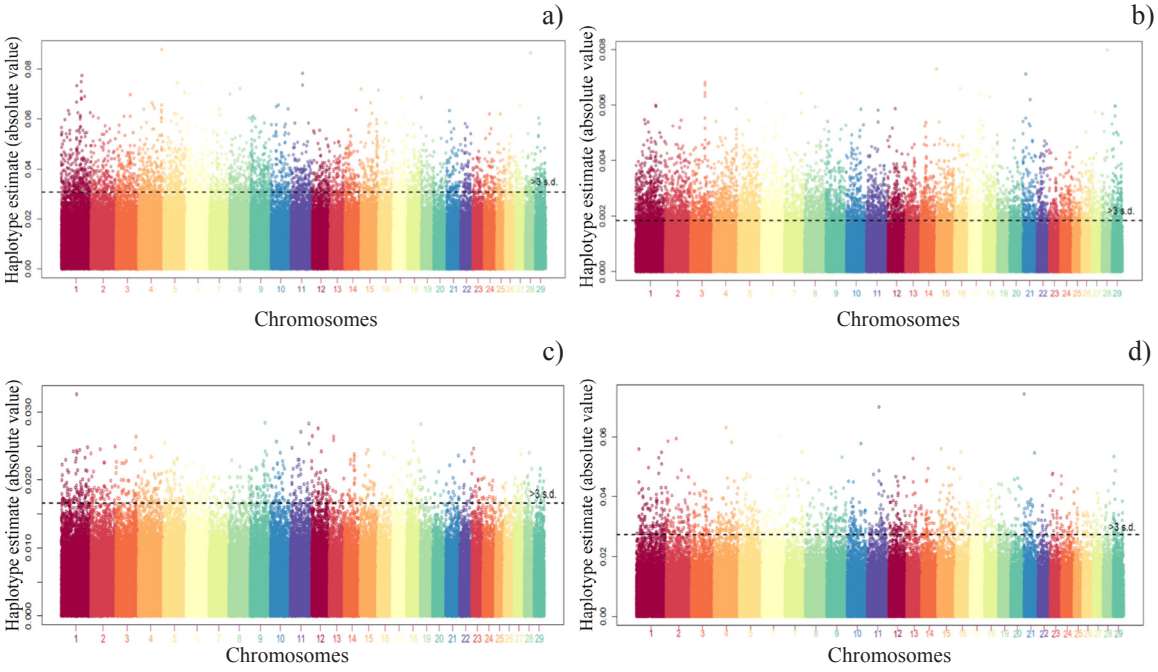

Figure 2. Manhattan plots with estimated haplotypes effects for a) PROT (kg of protein), b) GTI (Global Type Index), c) SCS (Somatic Cell Score) and d) DO (Days Open) traits.

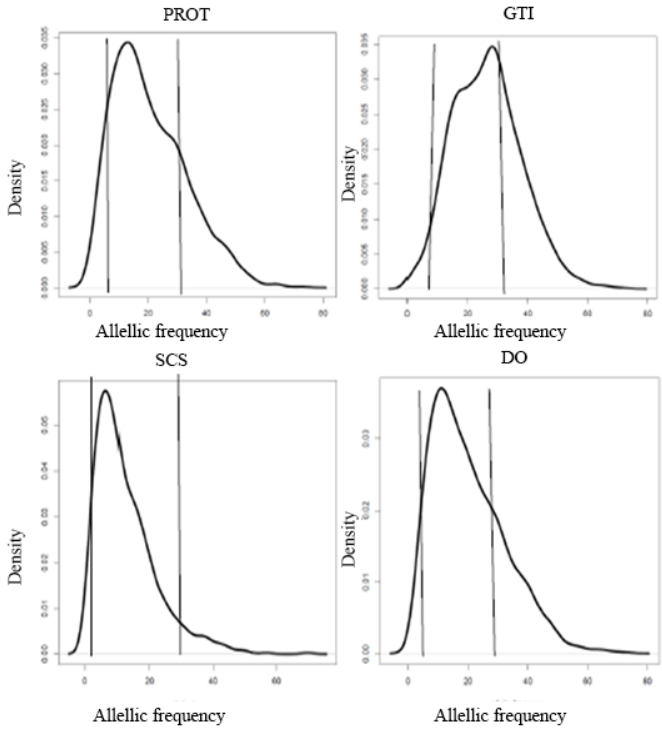

Figure 3. Allele frequency distribution for haplotypes with effect estimate exceeding the 3 SD threshold. PROT (kg of protein); GTI (Global Type Index); DO (Days Open); SCS (Somatic Cell Score)

model. This reduction was lower for PROT (15%), with no clear trend associated to the number of haplotypes included. However, the reduction for residual variance for GTI, DO and SCS was relevant when sequence haplotypes were taking into account (62-89%).

Discussion

Haplotypes have been extensively explored in human genetics research (Curtis *et al.*, 2001; Gabriel *et al.*, 2002; Chapman *et al.*, 2003; Curtis, 2007). More recent studies in animal breeding have explored the use of haplotypes for genomic prediction of breeding values, but using low to medium density marker data (Calus *et al.*, 2008, 2009; Villumsen *et al.*, 2009; Boichard *et al.*, 2012; Schrooten *et al.*, 2013). Using haplotypes in genome-enhanced prediction is a reasonable approach, under the hypothesis that haplotypes are expected to be in stronger LD to the causative mutations (or QTLs) than any single marker (Hyten *et al.*, 2007; Hamblin & Jannink, 2011). In our case, haplotypes were extracted by version 3 of findhap. This program was designed for SNPs chips and the versions have been improved and could be used for sequence data, which we have been proven in this study. More details on findhap program can be found in VanRaden *et al.* (2011).

This study detected long haplotypes, which supports the existence of long-range LD in the Holstein breed and validates the interest of haplotype analysis. Furthermore, when it comes to sequence data, haplotypes offer the possibility to reduce the number of explanatory variables in genomic prediction models compared with the individual SNP approaches. Several methods have been used to construct haplotypes for genomic evaluation (Calus *et al.*, 2008, 2009; Boichard *et al.*, 2012; Cuyabano *et al.*, 2014). The construction of haplotypes at a particular SNP position by merging this SNP with the flanking markers is straightforward. However, due to the short distance between the markers, most of the resulting haplotypes included a small number of over-represented alleles together with a large number of alleles with low frequencies within the population (Jónás *et al.*, 2016). It is expected that

Table 3. Proportion of genomic breeding values variance captured by sequence haplotypes for each trait.

	PROT (%)	GTI (%)	DO (%)	SCS (%)
All	32.75	71.93	73.76	99.90
>1SD BL	10.92	NC	53.93	33.30
>3SD BL	1.06	5.24	11.64	15.29

PROT (kg of protein); GTI (Global Type Index); DO (Days Open); SCS (Somatic Cell Score); BL (Bayesian LASSO); NC (no convergence obtained)

Table 4. Polygenic (σ_g^2) and residual (σ_e^2) variance for the traits analysed with or without including haplotypes.

	PROT		GTI		DO		SCS	
	σ_g^2	σ_e^2	σ_g^2	σ_e^2	σ_g^2	σ_e^2	σ_g^2	σ_e^2
>1SD BL	253.27	15.02 (6.3)	NC	NC	21.70	13.54 (8.6)	22.47	6.87 (2.5)
>3SD BL	376.70	11.28 (5.5)	0.88	0.073 (0.03)	81.16	7.34 (3.1)	36.58	13.77 (3.9)
Only GRM	429.28	13.29 (5.8)	1.23	0.67 (0.02)	127.09	19.17 (5.0)	54.41	27.51 (4.8)

PROT (kg of protein); GTI (Global Type Index); DO (Days Open); SCS (Somatic Cell Score); BL (Bayesian LASSO); NC (no convergence obtained); GRM (Genomic Relationship Matrix).

there is an optimal haplotype length, which depends on the distance between the markers and extent of LD in the population. Reliabilities for GEBV were investigated by simulation to test the hypothesis that there is an optimal haplotype size for genomic predictions; however, studies on real data in dairy cattle are limited. Villumsen *et al.* (2009) hypothesized that there is an optimal haplotype size for genomic predictions, showing a clear relationship between the size of haplotypes used in the prediction model and the reliabilities obtained with 30k SNP chips. In general they found high reliabilities for all the tested haplotypes lengths. Therefore, the performance of large haplotypes may become poor because the number of haplotype 'alleles' increases quickly with increasing haplotype size. This generates fewer unknowns to estimate. The optimal size of haplotypes was 10 SNP for heritabilities of 0.30 and 0.02 for the haplotype lengths they tested, because these haplotype sizes gave the highest reliabilities.

The optimal haplotype size is very dependent on marker spacing and marker frequencies. If marker distance is low, the nearest marker may not be the best predictor of the QTL effect, and a better predictor may be found at larger distances (Zondervan & Cardon, 2004), depending largely on the allele frequency of both causal mutation and the SNP variant. On the other side, a recent study showed that more accurate predictions can be obtained with haploblocks with a predetermined number of SNPs (Boichard *et al.*, 2012). There are previous studies using haplotype blocks in cattle, but with different parameters, such as breed, type of markers, marker density, or genome location. These studies yielded average haplotype block sizes ranging from a few kilobases [*e.g.* 5.7 kb considering 2 or more SNPs (Villa-Angulo *et al.*, 2009), 26.2 kb considering 4 or more SNPs (Kim & Kirkpatrick, 2009)] to hundreds of kilobases (*e.g.* 700 kb, Gautier *et al.*, 2007), but they used smaller marker densities than in our study, with an average distance of 62 kb between adjacent markers (Qanbari *et al.*, 2010).

This study suggests that haplotypes from sequence data may provide valuable information on genetic variance and may lead to the development of more efficient strategies to identify genetic variants associated with traits of economic interest. Sequence haplotypes captured part of the polygenic effect, but they also contributed to reduce the residual variance, suggesting that they can account for additional variance that is not captured by the polygenic effect. This was especially relevant for DO and SCS, which support the hypothesis that sequencing information can contribute further to explain the statistical genetic architecture of these traits undergoing lower selection pressure. Genomic predictions using a set of appropriately selected haploblocks are expected to achieve higher prediction accuracy while reducing the amount of predictor variables in prediction models. Using preselected haploblocks for genomic prediction is an important area of future research, as they are expected to increase the GEBV reliability as well as its persistency across generations. This seems even more relevant on traits that have undergone lower selection intensities, such as fertility or disease resistance. Besides, computation time can be considerably reduced compared to models using SNPs as in whole-genome sequence data. This study allowed us to observe the possibilities of incorporating sequenced data from the 1000 bull genomes project in routine genomic evaluations. Some previous studies supported the hypothesis that the use of sequence data would result in a larger predictive accuracy in genome-assisted evaluations (Meuwissen & Goddard, 2010; MacLeod *et al.*, 2013), but they also highlighted the need of either increasing the number of individuals in the training dataset or pre-selecting SNPs based on other sources of information (*e.g.* Wimmer *et al.*, 2013; Hayes *et al.*, 2014). Hayes *et al.* (2014) obtained a very small increase of 2 % in prediction reliability using 1.7 million imputed sequence data compared to BovineHD chip genotypes.

Whole genome sequence data allows to include rare variants in genomic prediction and GWAS, which may explain some extra variation in the targeted complex traits and our results support the hypothesis in Gonzalez-Recio *et al.* (2015) who suggested that traits undergoing lower selection pressure would benefit more from next generation sequencing data. SNP arrays have limited power to capture such a variation, as the

SNP on these arrays are selected to have high MAF, and are therefore unlikely to be in high LD with the rare variants (Hayes *et al.*, 2015). However, the present study does not include rare variants as differentiating them from sequencing errors is not straightforward (Gonzalez-Recio *et al.*, 2015). Another important advantage of haplotypes over single SNP markers is their better ability to identify mutations in more than one loci. According to Curtis *et al.* (2001), allele frequencies may change very little when a mutation occurs at a locus, but the frequencies of variants in a haplotype are more likely to change when mutations occur in one or more loci of a haploblock Therefore, haplotypes may be better able to identify a QTL region than individual SNPs.

One limitation of this study is the reduced number of individuals (361) with phenotypic data that were used to estimate the effects of over 46,000 haplotypes when filtering on the 3 SD criterion. Thus, the number of haplotypes (p) was much larger than the number of observation (n). In this scenario, the QTL effect might be estimated with large error, which reduces the advantage of using sequence data compared to SNP genotypes for genomic prediction. The choice of the prior distribution for λ^2 could potentially influence the results. Consistent results and convergence were observed when using different scale hyperparameters between 0.0001 and 0.00001. These hyperparameters affected the convergence of the Monte Carlo Markov Chain and should be chosen carefully, for example with a grid search, as done in this study with values ranging from 0.0000001 to 1.

In conclusion, the algorithm implemented in findhap can extract haplotypes to be used in genome-enhanced evaluations, although parameters must be tuned carefully for an efficient implementation, applying prior biological knowledge to approximate an appropriate length of the haplobocks. Given the availability of data from the 1000 bull genomes project, haplotypes with a larger effect would capture a small proportion of genetic variance, but they contribute to explain additional genetic variance mainly for traits that have not undergone high selection intensity such as fertility or health traits. The increase in predictive accuracy must be checked in future studies using imputation on the whole genotyped population and through cross-validation. Also, further research is needed to include low and rare variants in the statistical models.

Acknowledgments

The authors thank CONAFE for the phenotypic data provided and Drs Ana Fernández, Maria Saura, Beatriz Villanueva, Raquel de Paz and Almudena Fernández for the preparation of sequence data.

References

Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, Druet T, Genestout L, Colleau JJ, Journaux L, *et al.*, 2012. Genomic selection in French dairy cattle. Anim Prod Sci 52: 115-120. https://doi.org/10.1071/AN11119

Calus MPL, Meuwissen THE, De Roos APW, Veerkamp RF, 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553-561. https://doi.org/10.1534/genetics.107.080838

Calus MPL, Meuwissen THE, Windig JJ, Knol EF, Schrooten C, Vereijken ALJ, Veerkamp RF, 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genet Sel Evol 41: 11. https://doi.org/10.1186/1297-9686-41-11

Chapman JM, Cooper JD, Todd JA, Clayton DG, 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56: 18-31. https://doi.org/10.1159/000073729

Curtis D, 2007. Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association. BMC Genet 8: 49. https://doi.org/10.1186/1471-2156-8-49

Curtis D, North BV, Sham PC, 2001. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. Ann Hum Genet 65: 95-107. https://doi.org/10.1046/j.1469-1809.2001.6510095.x

Cuyabano BC, Su G, Lund MS, 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genom 15: 1171. https://doi.org/10.1186/1471-2164-15-1171

Cuyabano BC, Su G, Lund MS, 2015. Selection of haplotype variables from a high-density marker map for genomic prediction. Genet Sel Evol 47: 61. https://doi.org/10.1186/s12711-015-0143-3

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C., *et al.*, 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46: 858-865. https://doi.org/10.1038/ng.3034

De Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D, 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9 (7): e1003608 https://doi.org/10.1371/journal.pgen.1003608

Druet T, Macleod IM, Hayes BJ, 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112: 39-47. https://doi.org/10.1038/hdy.2013.13

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al.*, 2002. The structure of haplotype blocks in the human genome. Science 296: 2225-9. https://doi.org/10.1126/science.1069424

Gautier M, Faraut T, Moazami-Goudarzi K, Navratil V, Foglio M, Grohs C, Boland A, Garnier JG, Boichard D, Lathrop GM, Gut IG, Eggen A, 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059-1070. https://doi.org/10.1534/genetics.107.075804

Gonzalez-Recio O, Daetwyler HD, MacLeod IM, Pryce JE, Bowman PJ, Hayes BJ, Goddard ME, 2015. Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in cattle. PloS one 10: e0143945. https://doi.org/10.1371/journal.pone.0143945

Hamblin MT, Jannink JL, 2011. Factors affecting the power of haplotype markers in association studies. The Plant Genome Journal 4: 145. https://doi.org/10.3835/plantgenome2011.03.0008

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME, 2009. Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92: 433-443. https://doi.org/10.3168/jds.2008-1646

Hayes BJ, Bowman PJ, Daetwyler HD, Goddard ME, 2015. Why can we impute some rare sequence variants and not others? Proc Assoc Advmt Breed Genet 21: 41-44.

Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, vander Jagt CJ, Capitan A, Pausch H, Stothard P, Liao X, *et al.*, 2014. Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. Proc 10th World Congr Genet Appl Livest Prod. Am Soc Anim Sci, Champaign, IL, USA.

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB, 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175: 1937-1944. https://doi.org/10.1534/genetics.106.069740

Ibañez-Escriche N, Gonzalez-Recio O, 2011. Review: Promises, pitfalls and challenges of genomic selection in breeding programs. Span J Agric Res 9: 404-413. https://doi.org/10.5424/sjar/20110902-447-10

Jónás D, Ducrocq V, Fouilloux MN, Croiseau P, 2016. Alternative haplotype construction methods for genomic evaluation. J Dairy Sci 99: 1-10. https://doi.org/10.3168/jds.2015-10433

Kim ES, Kirkpatrick BW, 2009. Linkage disequilibrium in the North American Holstein population. Anim Genet 40: 279-88. https://doi.org/10.1111/j.1365-2052.2008.01831.x

MacLeod I, Hayes B, Goddard M, 2013. will sequence snp data improve the accuracy of genomic prediction in the presence of long term selection? Proc Assoc Advmt Anim Breed Genet. http://www.aaabg.org/aaabghome/AAABG20papers/macleod20215.pdf

Meuwissen T, Goddard M, 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185: 623-631. https://doi.org/10.1534/genetics.110.116590

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H, 2010. The pattern of linkage disequilibrium in German Holstein cattle. Anim Genet 41: 346-356.

Schrooten C, Schopen G, Parker A, Medley A, Beatson P, 2013. Across-breed genomic evaluation based on bovine high density genotypes, and phenotypes of bulls and cows. Proc Assoc Advmt Anim Breed Genet, pp: 138-141

Van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, Veerkamp RF, 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol 46: 41. https://doi.org/10.1186/1297-9686-46-41

VanRaden PM, 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91: 4414-23. https://doi.org/10.3168/jds.2007-0980

VanRaden PM, O'Connell JR, Wiggans GR, Weigel K, 2011. Genomic evaluations with many more genotypes. Genet Sel Evol 43: 10. https://doi.org/10.1186/1297-9686-43-10

Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ, 2009. High-resolution haplotype block structure in the cattle genome. BMC Genet 10: 19. https://doi.org/10.1186/1471-2156-10-19

Villumsen TM, Janss L, Lund MS, 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. J Anim Breed Genet 126: 3-13. https://doi.org/10.1111/j.1439-0388.2008.00747.x

Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schön CC, 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195: 573-587. https://doi.org/10.1534/genetics.113.150078

Zondervan KT, Cardon LR, 2004. The complex interplay among factors that influence allelic association. Nat Rev Genet 5: 89-100. https://doi.org/10.1038/nrg1270