



**HAL**  
open science

## Selecting sequence variants to improve genomic predictions for dairy cattle

Paul M. Vanraden, Melvin E. Tooker, Jeffrey R. O'connell, John B. Cole,  
Derek M. Bickhart

► **To cite this version:**

Paul M. Vanraden, Melvin E. Tooker, Jeffrey R. O'connell, John B. Cole, Derek M. Bickhart. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genetics Selection Evolution*, 2017, 49 (1), pp.32. 10.1186/s12711-017-0307-4 . hal-01484834

**HAL Id: hal-01484834**

**<https://hal.science/hal-01484834>**

Submitted on 7 Mar 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH ARTICLE

Open Access



# Selecting sequence variants to improve genomic predictions for dairy cattle

Paul M. VanRaden<sup>1\*</sup>, Melvin E. Tooker<sup>1</sup>, Jeffrey R. O'Connell<sup>2</sup>, John B. Cole<sup>1</sup> and Derek M. Bickhart<sup>1</sup>

## Abstract

**Background:** Millions of genetic variants have been identified by population-scale sequencing projects, but subsets of these variants are needed for routine genomic predictions or genotyping arrays. Methods for selecting sequence variants were compared using simulated sequence genotypes and real July 2015 data from the 1000 Bull Genomes Project.

**Methods:** Candidate sequence variants for 444 Holstein animals were combined with high-density (HD) imputed genotypes for 26,970 progeny-tested Holstein bulls. Test 1 included single nucleotide polymorphisms (SNPs) for 481,904 candidate sequence variants. Test 2 also included 249,966 insertions-deletions (InDels). After merging sequence variants with 312,614 HD SNPs and editing steps, Tests 1 and 2 included 762,588 and 1,003,453 variants, respectively. Imputation quality from findhap software was assessed with 404 of the sequenced animals in the reference population and 40 randomly chosen animals for validation. Their sequence genotypes were reduced to the subset of genotypes that were in common with HD genotypes and then imputed back to sequence. Predictions were tested for 33 traits using 2015 data of 3983 US validation bulls with daughters that were first phenotyped after August 2011.

**Results:** The average percentage of correctly imputed variants across all chromosomes was 97.2 for Test 1 and 97.0 for Test 2. Total time required to prepare, edit, impute, and estimate the effects of sequence variants for 27,235 bulls was about 1 week using less than 33 threads. Many sequence variants had larger estimated effects than nearby HD SNPs, but prediction reliability improved only by 0.6 percentage points in Test 1 when sequence SNPs were added to HD SNPs and by 0.4 percentage points in Test 2 when sequence SNPs and InDels were included. However, selecting the 16,648 candidate SNPs with the largest estimated effects and adding them to the 60,671 SNPs used in routine evaluations improved reliabilities by 2.7 percentage points.

**Conclusions:** Reliabilities for genomic predictions improved when selected sequence variants were added; gains were similar for simulated and real data for the same population, and larger than previous gains obtained by adding HD SNPs. With many genotyped animals, many data sources, and millions of variants, computing strategies must efficiently balance costs of imputation, selection, and prediction to obtain subsets of markers that provide the highest accuracy.

## Background

Accuracy of genomic predictions can be improved by using more variants, including variants that are pre-selected for their effect, located near genes or within genes, predicted to affect gene function, or known to be

causal. Past analyses often gave equal weight to evenly spaced markers, whereas new analyses can focus on potential quantitative trait loci (QTL) or preselected variants that are more closely linked to QTL. Nearly 40 million variants have been identified from whole-genome sequence (WGS) data for over 1500 bulls, and several strategies to impute these variants to additional animals and use them in genetic evaluation for economic traits show potential [1–8]. For example, candidate variants

\*Correspondence: Paul.VanRaden@ars.usda.gov

<sup>1</sup> Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA

Full list of author information is available at the end of the article

can be targeted to specific traits such as genes related to fertility, thereby slightly improving reliability for daughter pregnancy rate by 0.2 percentage points when 39 single nucleotide polymorphisms (SNPs) were added to the marker set used for genomic prediction [9]. The number of sequenced animals should continue to increase as researchers examine more families and the costs of generating data continue to decrease.

Imputing, selecting, and predicting effects for millions of variants and many thousands of individuals require efficient computation. Computational costs, which are proportional to the number of variants multiplied by the number of individuals, could exceed the marginal benefits from adding more variants. Variants within or near genes should improve the reliability of predictions, and direct use of causal variants is preferred to using linked markers. Strategies to choose variants for inclusion on genotyping arrays of different densities or in routine predictions were developed and compared using simulated data for Holstein bulls. Here, we first examined simulated data and then real sequence genotypes from the 1000 Bull Genomes Project [10].

The goals of this study were to (1) compare the reliability of prediction from sequence, array, and combined data as well as different types of variants, (2) test the methods first on simulated data before applying them to real sequence data imputed for a large reference population, and (3) investigate editing, imputation, and computing strategies that are efficient for even larger genotyped populations.

## Methods

### Simulated sequence data

Our simulation was designed to closely mimic an actual large-scale sequencing project for cattle, in which a subset of ancestor bulls had WGS data, another subset of ancestor bulls had high-density (HD) SNP-array genotypes, and most bulls had medium-density genotypes. Sequence variants were simulated for 26,984 Holstein bulls in the US reference population in December 2014 using a pedigree file that included 112,905 animals, and the sequences were then reduced to mimic the actual available array genotypes. Among these animals, the 1000 bulls that had the most daughters had genotypes observed for 30 million sequence variants, and 773, 24,863 and 343 other bulls that had fewer daughters were genotyped with 600,000 (600 k), 60,000 (60 k) and 12,000 SNPs, respectively. Each simulated chip was an evenly spaced subset of the previous chip and the sequence variants. The 30 million variant sites were randomly located across 30 chromosomes each 100 million bases long, and all variants had two alleles. The genotypes were simulated using *genosim* software [11], which generates founding

chromosomes with linkage disequilibrium (LD) and descendant chromosomes with recombination using the actual pedigree of the bulls. A parameter of 0.9998 was selected to generate average LD similar to that in the real sequence dataset, as in previous tests [12].

Editing reduced the list of variants to 8.4 million by removing SNPs with a minor allele frequency (MAF) lower than 0.01 and a level of LD less than 0.95 with any remaining neighbouring SNP, but all 0.5 million variants that were within or near the 10,000 (10 k) simulated QTL were retained. The QTL were located randomly across the genome, and the 25 variants on either side were retained. No actual genes were simulated, only the QTL and other variants. If any of the 350 variants on either side of a specific marker were correlated i.e. with an  $|r|$  higher than 0.95, editing based on LD retained one variant and removed all others that had an  $|r|$  higher than 0.95 with that variant. The 600 k SNPs were all retained to improve imputation, and the 505,210 SNPs that were within 2500 bases of a true QTL were retained to mimic bioinformatic selection using gene positions. The selected SNPs were imputed for all bulls. Strategies were compared to choose the most significant variants or those with the largest estimated variances or effect sizes for five independent traits using individual regressions on each variant or multiple regression on all variants.

Breeding values for five independent traits were simulated by summing across effects of the 10 k QTL. The five traits were not true replicates because the QTL locations did not vary, only the effects, mimicking quantitative inheritance where each QTL may affect most traits very little but some traits more. A heavy-tailed distribution was generated from normally-distributed effects ( $q$ ) raised to the power of  $2.7^{|q|-2}$  such that the largest effect contributed 3 to 13% of the genetic variance, the largest 10 effects contributed 20 to 34%, the largest 100 contributed 57 to 63%, and the largest 1000 contributed 90 to 93%. Actual traits may be controlled by QTL with smaller effects, however, most actual traits had at least one QTL as large as those simulated here [13]. Simulated phenotypes for five independent traits had reliabilities equal to those for milk evaluations of the actual bulls.

After imputing the 8.4 million edited variants for all bulls, the variants with the largest effect estimated by genomic prediction or the most significant variants from genome-wide association (GWA) analysis were selected. The oldest 17,896 bulls were used as the reference population, and true breeding values (TBV) of the 9088 younger bulls were used to validate predictions from the selected variants. In all tests, the phenotypes used for estimating effects and selecting variants were only from the truncated reference data so that validation phenotypes were independent and tests should be unbiased.

Many of the reference bulls and a few validation bulls had sequence data included in the 1000 Bull Genomes Project and used for variant discovery, which might bias estimates of allele frequency, but should not bias the phenotypic effects.

Variants can be selected based on the highest significance test, largest absolute effect, or largest genetic variance contributed by the locus, which is computed as  $2p(1-p)\alpha^2$ , where  $p$  is the allele frequency and  $\alpha$  is the allele substitution effect. All three methods were compared. Selecting the variants that contribute the most variance has more theoretical appeal and results in variants with higher MAF, which could also contribute to improve imputation accuracy. Using the nonlinear Bayes A algorithm of VanRaden et al. [12], the highest ranking markers were selected based on their largest effect or largest variance regardless of their location. Using GWA, the significance of each variant was tested conditional on neighbouring variants already included, and the tests were then combined for each of the five independent traits into an overall measure of significance. The single regression model in GWA was processed using MMAP [14, 15] and included SNP as a fixed effect and breeding values as random effects modelled with pedigree relationships. Pedigree information was used rather than genomic relationships based on sequence data to separate the individual effect of SNPs from the random, polygenic effect. Multiple regression requires hundreds of iterations to converge, whereas GWA can test many variants without iteration.

Genomic predictions from 60 or 600 k SNPs were compared with predictions from additional markers selected also using Bayes A multiple regression. To mimic the selection process used to design the GeneSeek HD version 1 chip [16], the top 5000 HD SNPs for each of the five traits were selected, and the combined set of 23,600 (24 k) selected SNPs after removing 1400 duplicates were added to the 60 k SNPs. To mimic selection on net merit [17], another test selected 24 k SNPs with the largest variance averaged across the five traits instead of selecting the top SNPs for each trait and then combining them.

Selection based on sequence variants should improve accuracy more than selection on HD SNPs, but the previously genotyped SNPs must be retained during imputation because sequence variants are not available for most animals. Genomic predictions included the 600 k SNPs plus 500 k sequence variants near QTL totalling 1.1 million variants, which was similar to the analysis of real data by Hayes et al. [10]. The variants that were chosen in close proximity to QTL are referred to as the genic subset of WGS variants although gene locations were not simulated, only QTL locations were. Final tests of the simulated data added the 10 k true QTL to the 60 k SNPs,

and an upper limit on reliability was obtained using only the imputed QTL in prediction with no prior variance assigned to the markers, the parameter of the heavy-tailed distribution set to the true parameter, and polygenic variance set to 0% instead of the 10% in other tests.

#### Real variants derived from population-scale WGS data

SNP and insertion-deletion (InDel) calls (sequence variants) from run 5 of the 1000 Bull Genomes Project [18] were released in July 2015. Sequence variants for 444 Holstein animals and HD imputed genotypes for 26,970 progeny-tested Holstein bulls were combined by imputation using findhap software (version 3) [19]. The total number of variants identified in run 5 was equal to 38 million SNPs and 1.7 million InDels, but many of those variants are monomorphic within the Holstein breed. InDels were on average 3 bp long and no more than 86 bp. Imputed sequence genotypes from the 1000 Bull Genomes Project data were set to missing if none of the three genotype probabilities (AA, AB, or BB) were higher than 0.98 as estimated by Beagle [20].

The HD genotypes of 2394 Holsteins mainly from North America, Italy, and Great Britain were used to impute genotypes of 590,363 other Holsteins that had genotypes obtained mainly by using SNP chips with 50,000 or fewer SNPs. The imputed HD genotypes of bulls used in this study were a subset of those animals. The original 777,962 HD SNPs were reduced to 312,614 by removing highly correlated markers with an  $|r|$  higher than 0.95 and by further editing before imputation with findhap (version 3) [12]. To verify direction and consistency of allele codes, genotypes called from sequences were matched to corresponding chip SNPs for 179 Holstein or red Holstein animals that had SNP genotypes imputed in the US database and sequences in the 1000 Bull Genomes Project database.

Variants with a MAF lower than 0.01, incorrect map locations, an excess of heterozygotes, or low correlations ( $|r| < 0.95$ ) between sequence and HD genotypes for the same variant were removed. A few hundred sequence variants were removed in specific regions that were known to be mapped incorrectly in the UMD3.1 bovine reference assembly. Most map issues had been previously detected by using small sets of SNPs that were lowly correlated to adjacent sets within windows that had excessive total numbers of haplotypes [12].

After merging sequence and HD data, Mendelian conflicts between parents and progeny were set to missing for 0.01% of the genotypes. The percentage of conflicts was expected to be small because both the HD and sequence genotypes had been previously edited. About 1% of the HD imputed genotypes were unknown in the findhap output, and their allele frequencies were

substituted when used in genomic prediction. All HD SNPs that were also in the sequence data were retained except in cases when the absolute correlation between HD SNPs was lower than 0.95. This editing step removed less than 1000 (0.3%) of the HD SNPs because a similar edit had previously been applied before imputation [12].

Three different sets of variants were imputed to test the use of candidate SNPs (Test 1), candidate SNPs and InDels (Test 2), and candidate SNPs, InDels, and intergenic and intronic variants (Test 3). Predictions and QTL discovery using Test 3 data will be reported separately. The initial edits for sequence genotypes used in Tests 1 and 2 were revised in Test 3 because imputation accuracy decreased when millions of intergenic and intronic variants were included. The new edits for Test 3 computed statistics across all samples to improve imputation accuracy instead of editing each animal individually. The VCF file contains three genotype probabilities from Beagle, and the editing done for Tests 1 and 2 simply retained any genotype with a probability higher than 0.98. The new edits were based on an individual probability higher than 0.95, and after processing all animals, a second edit deleted any variant that had more than 5% missing genotypes for low frequency variants ( $MAF < 0.10$ ) or more than  $MAF/2$  missing genotypes for more common variants. Thus, variants with  $MAF = 0.50$  were not used if more than 25% of the called genotypes had a probability below 95%. The third new edit for Test 3 checked for Hardy–Weinberg equilibrium and deleted variants that had 1.5 times more heterozygotes than the expected fraction of  $2p(1 - p)$ . After these edits, only 3,148,506 variants remained.

Quality and orientation of calls were examined using 179 bulls that had both sequence and HD genotypes. After reversing the orientation of the HD SNPs to match sequence data and keeping the sequence instead of the HD genotype for animals that had both, the two datasets were combined, resulting in 27,235 animals. Quality of imputation was assessed by keeping 404 of the sequenced animals in the reference population and randomly choosing 40 animals as a test set. Their sequence genotypes were reduced to the subset of genotypes that were in common with HD genotypes and then imputed back to sequence. The percentage of imputed genotypes that matched the original genotypes was the simple measure of sequence imputation accuracy.

Genomic predictions were computed using deregressed evaluations from August 2011 for 33 traits and 19,575 bulls. Predictions were tested using 2015 data of 3983 bulls with daughters that were first phenotyped after August 2011. Reliabilities were estimated from the squared correlations of predictions with the deregressed evaluations, divided by their reliabilities to account

for error variance, and adding the difference between observed and expected reliability of parent average to account for selection [21]. Regressions of 2015 data on 2011 predictions were compared to the expected value of 1.0.

Test 1 combined 481,904 candidate sequence SNPs with HD genotypes for 312,614 markers and a total of 762,588 variants. The candidate variants included 107,471 variants in exons, 9422 in splice sites, 35,242 in untranslated regions at the beginning and end of genes, 254,907 within a 2-kb region upstream, and 74,862 within a 1-kb region downstream, for a total of 481,904 candidate variants based on the Ensembl gene annotation [22] database version 79 released in 2015 ([ftp://ftp.ensembl.org/pub/release-79/gtf/bos\\_taurus](ftp://ftp.ensembl.org/pub/release-79/gtf/bos_taurus)). Test 2 also included any InDels that were located within genes or within the regions 2 kb upstream and 1 kb downstream of genes. Imputed data of Test 3 were used only for GWA because genomic predictions converged too slowly with more than 3 million variants, and the GWA results from real data will be reported separately. Additional file 1: Table S1 lists the variants included in each test.

A subset of variants was selected for potential use in routine genomic prediction by applying methods similar to those used previously to select the HD SNPs with the largest effects in the national evaluation [16] except that only Holstein data were used in the current test. The top 1000 SNPs by absolute effect size for each of the 33 traits were selected from Test 1 and merged to eliminate duplicates. These 16,648 sequence variants with the largest effects were selected from the analysis of 762,588 markers and added to the 60,671 markers used previously. However, 6584 or about 10% of those previously used markers were not called as variable and thus not reported in the sequence data and were not used in the final test set of 70,735 markers.

## Results

### Simulated sequence

Edits for MAF and LD removed 3.4 and 18.4 million variants, respectively, from the simulated WGS variants from our 1000 bull founder population, which reduced the variant list from 30 million initial simulated variants to 8.4 million that included the 600 k array SNPs and the 505,210 genic variants. For the 26.6 million variants with a MAF higher than 0.01, the maximum absolute correlation with any of the 350 variants on either side was on average equal to 0.96. High correlations improve imputation and also indicate that most QTL can be efficiently traced by nearby markers.

Average reliability of prediction was equal to 28.4% based on the simulated parent average, 77.8 and 80.1% based on the 60 and 600 k chips, respectively, 79.2%

based on the markers selected by GWA from the 600 k chip, and 87.2% based on only the 10 k imputed true QTL with no weight on the markers (Table 1). The reliability gain of 2.3 percentage points obtained for the 600 k compared with the 60 k SNPs is larger than reported earlier from either simulated (0.9) or actual (0.4) genomic predictions [12]. The previous results led to the conclusion that simply adding more markers resulted only in small improvements because prior variance for each marker was smaller, causing more shrinkage for all marker effects. Also, the additional markers were imputed rather than directly observed.

In Table 1, the other variant subsets were selected using effects from multiple regression instead of GWA. Adding 24 k SNPs from the 600 k with the largest effects to the 60 k SNPs resulted in higher reliability by 2.2 percentage points than adding 24 k SNPs selected by GWA and also in 1.3 percentage points higher than using all 600 k SNPs, which was consistent with previous results from real data [16]. Selecting SNPs on effect variance was expected to be more efficient than selecting on effect size, but effect size resulted in slightly higher reliability (81.5 vs. 81.2%). The increased MAF should have improved imputation accuracy, but only 19% of the SNPs differed between the two selection strategies. Selecting 24 k SNPs based on an average of the five traits to mimic index selection (results not shown in Table 1) led to about only 50% of the markers being in common with the other two strategies and resulted in slightly lower reliability than selecting for each trait and then combining them (81.1 vs. 81.2%).

The genic subset of 1.1 million simulated sequence variants resulted in a reliability of 86.4%, which was much higher than the 81.5% obtained from the best analysis from selecting 600 k SNPs and only about 1 percentage point less than the 87.2% maximum obtained by using just the 10 k true QTL. This confirms that selection of variants near genes improves accuracy if all genes are known and all variation is associated with genes, which is in agreement with Pérez-Enciso et al. [6]. Including 1.1 million

variants in routine evaluations or on chips is difficult, but 60 k SNPs plus the top 24 k SNPs that are chosen from the 1.1 million by multiple regression resulted in a reliability of 85.0%. If the 10 k true QTL were added to the 60 k SNPs but were not given extra prior variance, the reliability was then only 84.5% because too much prior variance was assigned to the 60 k SNPs compared to the 10 k QTL. All tests of simulated data had regressions of TBV on genomic predictions that averaged 1.02 to 1.05 across five traits, which is slightly higher than the expected value of 1.0; regressions on parent average averaged 0.98.

Computing resources are in Table 2 for each step run on an IBM X3850 X5 with 4 Intel X7560 CPUs (32 cores, 64 threads @ 2.27 GHz), and 512 GB of memory. Genotype simulation required 56 h with one thread and 210 GB of memory and the output was a 32-GB file. Calculation of linkage correlations between neighbouring sequence variants and pruning those that were highly correlated took 1 h with 10 threads and 27 GB of memory. Imputation of 8.4 million variants for 26,984 bulls required 38 h with 20 threads and 13 GB of memory and the output was a 220-GB file. Selection of variants by GWA required only 30 min with 30 threads and very little memory. Genomic prediction for 1.1 million variants and five traits required 22 h with five threads and 20 GB of memory. Thus, GWA was faster for selecting variants, but multiple regression selected marker sets that gave more reliable predictions.

**Real variants**

Edits for real as well as simulated sequence variants are documented in Table 3. In the real data, 20 million of the initial 39 million variants were removed because of low MAF, and another 13 million were removed because of high linkage with neighbouring variants. Further edits in Tests 1 and 2 retained only the HD markers, candidate SNPs, and candidate InDels. In Test 3, 3 million of the remaining variants with lower genotype probabilities were removed to improve imputation accuracy.

**Table 1 Reliabilities for five simulated traits from ten sources of genetic information**

Trait	Parent average	60 k	60 k + 24 k <sub>GWA</sub>	60 k + 24 k <sub>ES</sub>	60 k + 24 k <sub>EV</sub>	60 k + 24 k <sub>G</sub>	60 k + 10 k QTL	Only 10 k QTL	600 k	1.1 m genic
1	24.4	77.9	79.2	81.6	81.3	85.4	84.6	87.2	80.3	86.7
2	31.2	77.9	79.3	81.4	81.2	85.3	84.9	87.7	80.1	86.7
3	32.7	78.3	79.5	81.7	81.5	84.9	85.0	87.8	80.4	86.1
4	23.3	76.6	77.7	80.2	79.8	83.5	82.9	85.9	78.6	84.8
5	30.4	78.3	80.0	82.5	82.2	86.0	85.2	87.5	81.2	87.6
Average	28.4	77.8	79.2	81.5	81.2	85.0	84.5	87.2	80.1	86.4

Reliabilities expressed as percentages; 24 k markers selected from 600 k SNPs by GWA *p* value, multiple regression effect size (ES) or effect variance (EV); 24 k markers selected from sequence SNPs in or near genes (G) by effect size; 600 k markers plus 500 k SNPs in or near genes (1.1 m genic) by effect size

**Table 2 Computer resources needed to select markers from 30 million simulated variants**

Variant selection step	Number of threads	Computational time (h)	GB of memory	GB of disk space
Simulate 30 million	1	56	210	32
Prune linkage	10	1	27	10
Impute 8.4 million	20	38	13	220
Select 25,000	30	0.5	<1	<1
Predict 1 million	5	22	20	<1

1000 sequenced and 25,984 genotyped bulls

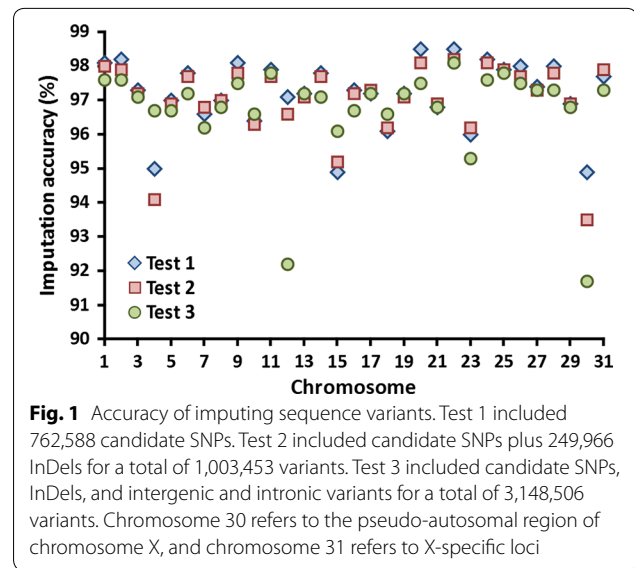
**Table 3 Edits applied to simulated data and real data from Test 3**

SNP edit	Simulated data	Real data
Original number of SNPs called	30 million	39 million
Removed for MAF of <0.01	3 million	20 million
Removed for linkage of >0.95	18 million	13 million
Removed for imputation inaccuracy	0	3 million
Remained after edits	8 million	3 million

Test 3 included candidate SNPs, InDels, and intergenic and intronic variants

Only 91% of the 60,671 chip SNPs currently used in official US evaluations were included in the sequence data. It is expected that some markers with a low MAF will be missing, but the average MAF of the 9% that were missing and the 91% that matched were both equal to about 0.28 in Holsteins. The missing markers are evenly scattered across the chromosomes; therefore, they probably do not indicate reference genome misassemblies but likely result from edits during variant identification [18]. The individual correlations of HD genotypes with sequence genotypes were mostly near +1 or -1, which indicates good quality for the 91% of HD SNPs present in the sequence data. About half of the genotypes had reversed allele coding compared to the sequence variant calls because sequence alternate alleles are coded as differences from a Hereford cow-derived reference genome, whereas the array alleles were in Illumina TOP encoding.

Average imputation accuracy was equal to 97.2% of correct genotypes for the 762,588 variants in Test 1 across all chromosomes, with a maximum of 98.5% for bovine chromosome BTA20 (BTA for *Bos taurus* chromosome) and BTA22 and a minimum of 94.9% for BTA15 and 95.0% for BTA4 (Fig. 1). The X chromosome was split into the pseudo-autosomal region, which was labelled as BTA30 with poor imputation and the X-specific loci labelled as BTA31; no Y loci were present. Imputation accuracy was equal to 97.0% for the 1,003,453 variants that included InDels in Test 2 and 96.7% for the 3,148,506 variants that also included intronic and intergenic variants in Test 3. The percentages are inflated because they



**Fig. 1** Accuracy of imputing sequence variants. Test 1 included 762,588 candidate SNPs. Test 2 included candidate SNPs plus 249,966 InDels for a total of 1,003,453 variants. Test 3 included candidate SNPs, InDels, and intergenic and intronic variants for a total of 3,148,506 variants. Chromosome 30 refers to the pseudo-autosomal region of chromosome X, and chromosome 31 refers to X-specific loci

include the HD SNPs that were already present. If HD SNPs are not counted, accuracies of 95.3, 95.6, and 96.4% for just the new variants were found for Tests 1, 2, and 3, respectively. The lower imputation accuracy for BTA12 in Test 3 was mainly caused by a gap between 72.4 and 75.2 Mb for which no SNPs were available on the HD array.

The total time required to prepare, edit, and impute the 762,588 variants for 27,235 animals ranged from 1 to 5 h per chromosome (Table 4) and was about 5 days for all 30 chromosomes. Data manipulation steps such as transposing the sequence data and merging with HD SNPs used one thread and took more time than the imputation, which used 20 threads and took less than 1 day.

Reliability of predictions improved by only 0.6 percentage points on average using 762,588 variants (481,904 candidate sequence variants and HD SNPs) compared with using HD SNPs only (Table 5). Inclusion of InDels decreased the advantage over HD SNPs to only 0.4 percentage points. Reliability improved by about 2.7 percentage points compared with 60 k SNPs only for the final set of 70,735 variants (60 k SNPs minus 6584 markers

**Table 4 Computation time<sup>a</sup> required with real sequence data for the longest (BTA1) and shortest (BTA29) chromosomes**

Computational step	BTA1	BTA29
Unzip VCF files	6	2
Read and transpose sequence	95	36
Subset sequenced animals	1	1
Subset matching HD markers	8	10
Merge sequence and HD data	143	6
Compute sequence linkage	3	1
Subset edited variants	3	1
Fix Mendelian conflicts	3	1
Impute with edited data	16	10
Reduce some sequence to HD data	1	1
Impute with reduced data	17	9
Total	296	78

<sup>a</sup> Time in minutes

that were not included in the sequence data plus 16,648 sequence variants with the largest effects). Reliability was equal to 35.2% based on parent average, 64.7% for predictions from 60 k SNPs only, 67.4% from 762,588 SNPs, 64.0% from HD SNPs only, 64.6% for HD plus genic SNPs, and 64.4% from HD plus genic SNPs and InDels. The 60 k SNPs already included the best SNPs selected from the HD chip [16], which may explain why 60 k predictions slightly outperformed HD predictions. For most traits, regressions of validation data on genomic predictions were near the expected value of 1.0 and changed little with the selected subset of variants used (Table 6).

For use with lower-density genotyping arrays, the list of 16,648 sequence variants was further restricted to 4822. Hand-made edits were applied to prevent too many candidate SNPs from all tagging the same QTL. Figure 2 provides an example for BTA5 of the SNPs that were retained or removed. The same list of 4822 SNPs was provided to Zoetis (Florham Park, NJ), GeneSeek (Lincoln, NE), and Genetic Visions (Middleton, WI) for potential inclusion on revised arrays. Benefits of adding the sequence SNPs directly to lower-density rather than only to medium- or higher-density arrays are that more young animals can be genotyped quickly and imputation loss can be avoided when including sequence SNPs in routine predictions. Re-genotyping or sequencing more reference animals could also help avoid imputation loss when estimating SNP effects for newly discovered variants.

## Discussion

### Comparison with previous studies

Previous studies used 5000 bulls with HD genotypes and 10 million variants from run 3 sequence data [8] or 4 million variants from run 4 [3], but sequence predictions in

those studies had slightly lower reliabilities than predictions from BovineHD or BovineSNP50 genotypes. The HD genotypes in those studies were all observed, but HD genotypes used in our study were mostly imputed. Use of run 3 or run 4 instead of run 5 sequence data could explain their slightly negative instead of slightly positive gains. The results from those studies and ours suggest that errors in the sequence data or remaining reference assembly mistakes that altered the order of variant sites could account for the small changes in prediction reliability when hundreds of thousands or millions of sequence variants were added.

Our results indicate that adding smaller numbers of selected sequence variants can be useful in routine prediction even if the analysis of all variants is not more accurate or feasible, which is consistent with previous conclusions for sequence [2] or HD data [16, 23]. Brøndum et al. [2] added 1623 sequence variants selected by GWA from multiple breeds to a custom chip and reported gains in reliability that averaged about 2 percentage points. Small improvements (0.2 percentage points) from adding SNPs that are located in genes associated with fertility were observed by Ortega et al. [9], which is consistent with gains reported in this and earlier studies [12]. Using selected sequence variants and giving extra weight to candidate variants or QTL can improve predictions across breeds [5, 24–27], but advantages of focusing on candidate variants decrease if not all QTL are in the variant set [6]. Multi-trait methods can detect QTL that single-trait methods might miss [28], and even uncorrelated traits can help separate QTL from markers if many independent traits are controlled by a limited number of QTL.

### Comparison of simulated and real selection

Properties of the real sequence data from the 1000 Bull Genomes Project were similar to those of the simulated data by VanRaden and O'Connell [29]. LD and MAF distributions in the real and simulated data are compared in Figs. 3 and 4, respectively. Overall, results were similar for real and simulated data, but more of the variants in the real data have a very low MAF or are in very low LD with neighbouring variants. Average MAF was the same (0.20) for the HD and genic SNPs in Test 1 but was lower (0.15) for the InDels added in Test 2, which could have affected imputation accuracy. Edits for MAF and for high LD reduced the 30 million simulated SNPs to 8.4 million, whereas the same edits reduced the 39 million real variants to 6.3 million (Table 3). Our edits for Test 3 were similar to those of Calus et al. [3], who obtained 4.1 million variants from Holstein data in run 4. The simulated variants were for one breed with a common pedigree, whereas the real variants were discovered in a wide



**Table 5 Reliability gains when adding real sequence variants to HD or 60 K**

Trait	Reliability for PA (%)	Gain for HD SNPs only	Gain for HD SNPs + 481,904 candidate SNPs <sup>a</sup>	Gain for HD SNPs + 481,904 candidate SNPs + indels	Gain for 60 k SNPs only	Gain for 60 k SNPs + 16,648 candidate SNPs <sup>b,c</sup>
Milk	37.9	34.1	33.9 (−0.2)	33.9	34.3	35.7 (1.4)
Fat	37.9	33.7	34.0 (0.3)	33.4	34.3	35.1 (0.8)
Protein	37.9	27.9	27.0 (−0.9)	26.7	27.5	28.2 (0.7)
Fat percentage	37.9	49.2	52.7 (3.5)	52.4	52.9	54.8 (1.9)
Protein percentage	37.9	42.1	41.6 (0.5)	43.0	41.6	44.3 (2.7)
Productive life	32.0	36.1	35.8 (−0.3)	36.4	35.6	38.2 (2.6)
Somatic cell score	34.7	35.9	36.1 (0.2)	37.1	35.1	37.0 (1.9)
Daughter pregnancy rate	31.5	30.8	30.0 (−0.8)	31.2	29.0	33.0 (4.0)
Cow conception rate	29.8	28.7	28.1 (−0.6)	28.8	28.9	31.8 (2.9)
Heifer conception rate	30.0	19.0	20.3 (1.3)	19.7	20.5	21.5 (1.0)
Sire calving ease	29.9	27.8	27.7 (−0.1)	25.2	24.5	28.5 (4.0)
Daughter calving ease	25.3	32.5	30.8 (−1.7)	29.9	31.5	31.4 (−0.1)
Sire stillbirth	29.0	7.6	7.3 (−0.3)	7.1	7.6	7.8 (0.2)
Daughter stillbirth	23.8	37.4	37.0 (−0.4)	35.8	35.4	38.0 (2.6)
Final score	36.2	24.7	25.5 (0.8)	25.8	24.6	27.8 (3.2)
Stature	38.2	30.4	32.4 (2.0)	32.8	30.3	34.7 (4.3)
Strength	37.4	29.9	31.8 (1.9)	31.8	29.9	34.5 (4.6)
Dairy form	37.4	33.8	35.3 (1.5)	35.8	35.0	38.2 (3.2)
Foot angle	36.7	17.3	17.6 (0.3)	18.2	17.2	19.6 (2.4)
Rear legs (side view)	37.3	21.9	22.7 (0.8)	22.0	22.1	24.1 (2.0)
Body depth	37.6	31.0	33.1 (2.1)	33.7	31.2	36.0 (4.8)
Rump angle	37.8	32.7	34.0 (1.3)	33.5	32.9	36.1 (3.2)
Rump width	37.1	29.2	30.4 (1.2)	30.2	29.1	32.5 (3.4)
Fore udder attachment	37.5	35.1	36.4 (1.3)	36.1	35.0	39.0 (4.0)
Rear udder height	37.3	24.7	25.7 (1.0)	25.8	24.1	27.3 (3.2)
Udder depth	38.0	40.2	42.6 (2.4)	42.8	40.6	44.6 (4.0)
Udder cleft	37.1	23.7	24.5 (0.8)	24.0	23.6	25.5 (1.9)
Front teat placement	37.6	32.6	33.4 (0.8)	32.3	30.9	35.0 (4.1)
Teat length	37.7	29.0	30.3 (1.3)	29.9	28.0	32.7 (4.7)
Rear legs (rear view)	36.0	20.7	20.3 (−0.4)	20.1	20.4	22.8 (2.4)
Feet and leg score	36.4	16.9	16.5 (−0.4)	16.6	15.9	18.3 (2.4)
Rear teat placement	37.4	33.1	33.6 (0.5)	32.1	32.9	35.2 (2.3)
Net merit	34.4	23.8	24.3 (0.5)	24.4	23.4	24.7 (1.3)
Average	35.2	28.8	29.4 (0.6)	29.2	29.5	32.2 (2.7)

Reliability gains in percentage points over parent average reliability

PA parent average

<sup>a</sup> Difference from reliability gain for HD SNPs only in parentheses

<sup>b</sup> Difference from reliability gain for 60 k SNPs only in parentheses

<sup>c</sup> Does not include 6584 60 k markers that were not available in sequence data

variety of breeds and only the variants that were polymorphic in Holsteins were retained. Also, the real data contain some false positive variants because of sequencing, alignment, and calling errors that are not modelled in the simulated data. Many variants had to be excluded from the Test 3 data because of low previous imputation

accuracy, whereas the simulated data was of high quality for all variants.

Gains in reliability from the use of real sequence data were smaller than from simulated data but higher than previous gains reported from HD data. Larger gains may be possible if the selected SNPs are added to arrays

**Table 6** Coefficients for regression of validation data on genomic predictions when adding real sequence variants to HD or 60 k

Trait	PA	HD SNPs only	HD SNPs + 481,904 candidate SNPs	HD SNPs + 481,904 candidate SNPs + InDels	60 k SNPs only	60 k SNPs + 16,648 candidate SNPs <sup>a</sup>
Milk	0.81	1.03	1.06	1.06	1.04	1.05
Fat	0.68	0.92	0.95	0.94	0.94	0.93
Protein	0.75	0.93	0.96	0.95	0.94	0.95
Fat percentage	0.97	1.14	1.13	1.12	1.12	1.09
Protein percentage	0.77	0.96	0.98	0.97	0.95	0.96
Productive life	1.24	1.30	1.32	1.25	1.27	1.25
Somatic cell score	0.89	1.09	1.10	1.06	1.08	1.06
Daughter pregnancy rate	1.20	1.47	1.49	1.48	1.43	1.43
Cow conception rate	0.72	0.94	0.95	0.92	0.91	0.91
Heifer conception rate	0.75	0.97	1.03	0.98	0.94	0.92
Sire calving ease	0.65	0.83	0.83	0.81	0.82	0.84
Daughter calving ease	0.80	1.04	1.03	1.02	1.04	1.02
Sire stillbirth	0.84	0.75	0.76	0.78	0.77	0.76
Daughter stillbirth	0.77	1.15	1.16	1.15	1.12	1.16
Final score	0.71	0.93	0.92	0.92	0.91	0.88
Stature	0.84	1.04	1.02	1.01	1.01	1.00
Strength	0.80	1.05	1.03	1.02	1.01	0.99
Dairy form	0.82	1.10	1.08	1.08	1.07	1.05
Foot angle	0.71	0.84	0.82	0.82	0.81	0.79
Rear legs (side view)	0.87	1.01	0.99	0.99	0.98	0.96
Body depth	0.76	1.01	0.99	0.99	0.97	0.96
Rump angle	0.80	1.08	1.07	1.05	1.06	1.05
Rump width	0.78	1.01	0.99	0.98	0.98	0.96
Fore udder attachment	0.80	1.06	1.04	1.03	1.03	1.01
Rear udder height	0.78	0.97	0.96	0.96	0.94	0.93
Udder depth	0.76	1.11	1.09	1.08	1.07	1.06
Udder cleft	0.87	1.00	0.99	0.99	0.98	0.95
Front teat placement	0.80	1.05	1.03	1.01	1.02	0.99
Teat length	0.91	1.06	1.06	1.04	1.04	1.03
Rear legs (rear view)	0.58	0.86	0.85	0.83	0.83	0.80
Feet and leg score	0.54	0.74	0.72	0.72	0.71	0.68
Rear teat placement	0.90	1.13	1.10	1.09	1.09	1.04
Net merit	0.85	0.82	0.84	0.81	0.83	0.81
Average	0.81	1.01	1.01	1.00	0.99	0.98

PA parent average

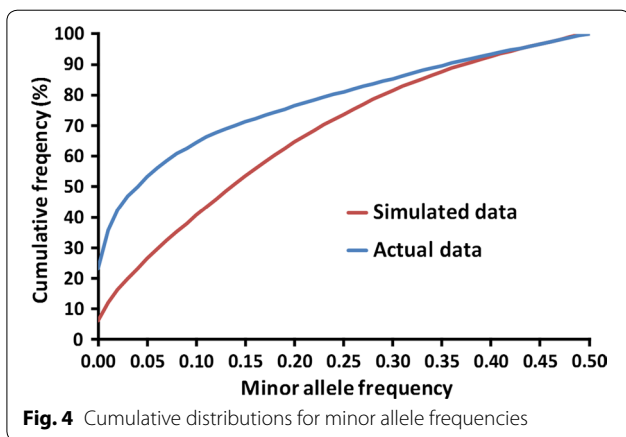
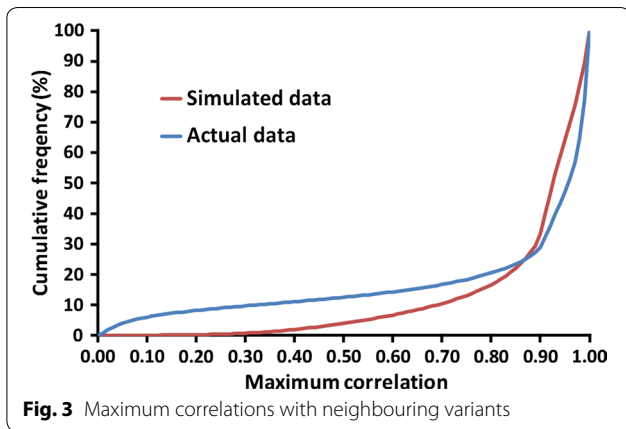
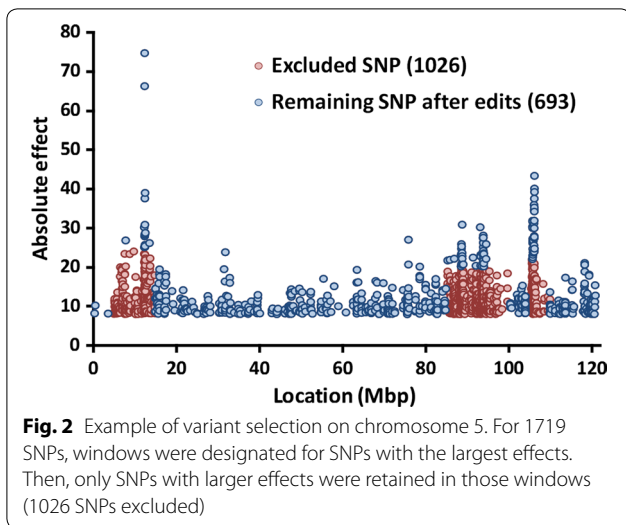
<sup>a</sup> Does not include 6584 60 k markers that were not available in sequence data

and genotyped directly with high accuracy instead of imputed from less accurate sequence data. Accuracies of genotypes from sequence variant calling can vary [30], whereas the error rate of Illumina BeadChip arrays is less than 1% for nearly all SNPs.

### Computation

Most computing steps in Table 4 were programmed in Fortran for efficiency, but several steps were in SAS for convenience. The SAS program used to merge sequence and HD data took only 6 min for the shortest

chromosome but 143 min for the longest one; it could be rewritten because it became a limiting step. Total times required for Tests 2 and 3 were only a little longer than those shown for Test 1 because imputation took a small fraction of the total time. Imputation of 8 million simulated variants took only 38 h with 20 threads for 25,984 reference bulls. Larger populations or variant sets can be imputed, but genomic predictions then become the limiting step. More research is needed on how to accurately and efficiently select the best subset of variants for routine use.



**Economic benefit**

Increasing the reliability of selection by 2.7 percentage points from 64.7 to 67.4% would add about \$3 million per year to national genetic progress. Additional

progress would be realized globally for foreign breeders that directly use the new genotyping arrays or that indirectly benefit by selecting breeding stock from the improved US population. Annual domestic progress is now about \$50 per cow and would increase to \$51 after multiplying by the accuracy ratio of 1.02, which is the square root of the reliability ratio (67.4/64.7). This higher accuracy has an annual national value of about \$3 million because each year 3.3 million of the 9.2 million US dairy cows are replaced. These annual gains are permanent and will accumulate. The initial cost of generating the US sequence data for the 88 dairy bulls contributed to the 1000 Bull Genomes Project was \$132,000 at current estimates of reagent costs (assuming a cost of ~\$1500 per sample). The return on investment from this research is high and greatly increased because of data sharing.

New animals will be directly genotyped for the selected variants and thus could have slightly higher reliability gains than in these tests that use imputed data, but most reference animals will still have imputed data. Re-genotyping old animals with the new arrays might be less expensive than additional sequencing to improve accuracy of imputation.

**Conclusions**

Variant selection is needed because routine genomic predictions cannot impute and include all of the millions of sequence variants for all animals. Significant gains in reliability are possible if the true QTL can be identified or if bioinformatic methods can choose the regions that are more likely to contain causative variants. Because individual QTL have such small effects, large reference populations are needed with phenotypes for the relevant traits and observed or imputed genotypes for the QTL or closely linked variants. Testing many individual traits gives more power because the effect of each QTL may be detectable only for a few traits, but these same QTL may have smaller effects on several correlated traits. Assigning more prior variance to the QTL or to the newly selected variants can improve reliability when estimating effects, but the SNPs from previous arrays must be retained during imputation because genotypes of previous animals include only the SNPs and not the new variants.

Computation becomes a limiting factor as reference populations and target populations grow in size. Total computing time was only a few days with up to 1000 sequences and 30,000 reference bulls, but more than 150,000 reference cows and 800,000 young animals were not included. Multiple regressions used for genomic prediction were more accurate than GWA for selecting variants but required much more computation time. Imputation allows

many more sequence variants to be tested, selected, and included in routine predictions to increase their reliability. For both the simulated and real data, gains from selecting and including candidate sequence variants were larger than from selecting HD SNPs.

## Additional file

**Additional file 1.** List of variants from Run 5 of the 1000 Bull Genomes Project used in Tests 1, 2 and 3. All edited variants are listed with 1 or 0 in the final three columns indicating if the variant was used in Test 1, 2 or 3. Fields included are VariantName, Chromosome, Location, VariantType, Test\_1, Test\_2, and Test\_3 in a space delimited file. Variant types use the 3-character codes EXN = exonic, SPL = splice site, UTR = untranslated region, UPS = upstream, DNS = downstream, SNP = other intronic or intergenic SNP, and IND = InDel.

## Authors' contributions

PMV, JRO, and DMB developed the experimental designs. MET and PMV performed many of the computations. PMV, JBC, and JRO drafted the paper. All authors read and approved the final manuscript.

## Author details

<sup>1</sup> Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA. <sup>2</sup> University of Maryland Baltimore, Baltimore, MD, USA.

## Acknowledgements

The authors thank Suzanne Hubbard for technical editing and manuscript improvement, George Liu and Steve Schroeder for assistance in generating the US sequence data, the 1000 Bull Genomes Project for global sequence data, the Council on Dairy Cattle Breeding for genotype, phenotype, and pedigree data, Interbull for global trait evaluations, and the anonymous reviewers for many helpful comments. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

Sequence data used in this study originated from the 1000 Bull Genomes Project [18]. WGS data of 234 individual bulls of the 1000 Bull Genomes Project are available at NCBI using SRA No. SRP039339 (<http://www.ncbi.nlm.nih.gov/bioproject/PRJNA238491>). Called genotypes for more recently sequenced bulls are available to participants in the international project. Pedigrees, phenotypes, and array genotypes used in this study can be obtained from the Council on Dairy Cattle Breeding (Bowie, MD) for research.

## Ethics approval and consent to participate

This study involved no animal experimentation and did not require any authorization from local ethics committee because DNA was extracted from commercialized semen straws.

## Funding

PMV, MET, DMB, and JBC were supported by appropriated project 8042-31000-101-00 (Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information) of the Agricultural Research Service, USDA. JRO was supported by a Specific Cooperative Agreement with the Agricultural Research Service, USDA.

Received: 27 September 2016 Accepted: 27 February 2017

Published online: 07 March 2017

## References

- Brøndum RF, Gulbrandsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics*. 2014;15:728.
- Brøndum RF, Su G, Janss L, Sahana G, Gulbrandsen B, Boichard D, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J Dairy Sci*. 2015;98:4107–16.
- Calus MPL, Bouwman AC, Schrooten C, Veerkamp RF. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. *Genet Sel Evol*. 2016;48:49.
- Druet T, MacLeod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* (Edinb). 2014;112:39–47.
- MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*. 2016;17:144.
- Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol*. 2015;47:43.
- van Binsbergen R, Bink MCAM, Calus MPL, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol*. 2014;46:41.
- van Binsbergen R, Calus MPL, Bink MCAM, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol*. 2015;47:71.
- Ortega MS, Denicol AC, Cole JB, Null DJ, Hansen PJ. Use of single nucleotide polymorphisms in candidate genes associated with daughter pregnancy rate for prediction of genetic merit for reproduction in Holstein cows. *Anim Genet*. 2016;47:288–97.
- Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, Vander Jagt CJ, et al. Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. In: Proceedings of the 10th world congress on genetics applied to livestock production: 17–22 August 2014; Vancouver. 2014. [https://asas.org/docs/default-source/wcgalp-proceedings-oral/183\\_paper\\_10441\\_manuscript\\_1644\\_0.pdf](https://asas.org/docs/default-source/wcgalp-proceedings-oral/183_paper_10441_manuscript_1644_0.pdf). Accessed 27 Dec 2016.
- VanRaden P, Sun C. genosim: Simulate genotypes, breeding values, and phenotypes; simulate DNA sequence read depth (numbers of A and B alleles); and resolve SNP conflicts between parent and offspring genotypes. In: Animal improvement program. Animal Genomics and Improvement Laboratory, ARS, USDA. 2014. <https://aipl.arsusda.gov/software/genosim>. Accessed 27 Dec 2016.
- VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. *J Dairy Sci*. 2013;96:668–78.
- Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, et al. Distribution and location of genetic effects for dairy traits. *J Dairy Sci*. 2009;92:2931–46.
- O'Connell J. MMAP: a comprehensive mixed model program for analysis of pedigree and population data. In: Proceedings of the 63rd annual meeting of the American society for human genetics: 22–26 October 2013; Boston. 2013. <http://www.ashg.org/2013meeting/abstracts/full-text/f130123097.htm>. Accessed 27 Dec 2016.
- O'Connell JR. MMAP user guide. 2016. <http://edn.som.umaryland.edu/mmap/index.php>. Accessed 27 Dec 2016.
- Wiggans GR, Cooper TA, VanRaden PM, Van Tassell CP, Bickhart DM, Sonstegard TS. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. *J Dairy Sci*. 2016;99:4504–11.
- VanRaden PM, Cole JB. AIP research report NM55: net merit as a measure of lifetime profit: 2014 revision. In: Animal improvement program. Animal Genomics and Improvement Laboratory, ARS, USDA. 2014. <https://aipl.arsusda.gov/reference/nmcalc-2014.htm>. Accessed 27 Dec 2016.
- Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet*. 2014;46:858–65.
- VanRaden PM. findhap.f90: Find haplotypes and impute genotypes using multiple chip sets and sequence data. In: Animal improvement program.

- Animal Genomics and Improvement Laboratory, ARS, USDA. 2016. <https://aipl.arsusda.gov/software/findhap>. Accessed 27 Dec 2016.
20. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81:1084–97.
  21. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci.* 2009;92:16–24.
  22. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. *Nucleic Acids Res.* 2015;43:D662–9.
  23. Saatchi M, Garrick DJ. Improving accuracies of genomic predictions by enriching 50K genotypes with markers from 770K genotypes at QTL regions. *J Dairy Sci.* 2014;97(Suppl 1):6.
  24. Kizilkaya K, Fernando RL, Garrick DJ. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J Anim Sci.* 2010;88:544–51.
  25. Iheshiolor OOM, Woolliams JA, Yu X, Wellmann R, Meuwissen THE. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet Sel Evol.* 2016;48:15.
  26. van den Berg I, Boichard D, Guldbbrandsen B, Lund MS. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. *G3 (Bethesda).* 2016;6:2553–61.
  27. van den Berg I, Boichard D, Lund MS. Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. *J Dairy Sci.* 2016;99:8932–45.
  28. Pausch H, Emmerling R, Schwarzenbacher H, Fries R. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. *Genet Sel Evol.* 2016;48:14.
  29. VanRaden PM, O'Connell JR. Strategies to choose from millions of imputed sequence variants. *Interbull Bull.* 2015;49:10–3.
  30. Baes CF, Bapst B, Seefried FR, Flury C, Signer-Hasler H, Garrick DJ, et al. Across-breed imputation with whole genome sequence data in dairy cattle. In: Proceedings of plant & animal genome XXIII: 10–14 January 2015; San Diego. 2015. <https://pag.confex.com/pag/xxiii/webprogram/Paper16562.html>. Accessed 27 Dec 2016.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

