Selection of Sequence Variants to Improve Genomic Predictions

J. R. O'Connell[1], M. E. Tooker[2], D. M. Bickhart[2], and P. M. VanRaden[2]
[1]University of Maryland School of Medicine, Baltimore, MD 21201, USA
[2]Animal Genomics and Improvement Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, MD 20705-2350, USA

Abstract

Methods of selecting sequence variants were compared using candidate sequence variants within or near genes for 444 Holsteins from run 5 (July 2015) of the 1000 Bull Genomes Project. Test 1 included single nucleotide polymorphisms (SNPs) for 481,904 candidate sequence variants within or near genes. Test 2 also included 249,966 insertions and deletions (indels). After merging sequence variants with 312,614 high-density (HD) SNPs and editing, Test 1 included 762,588 variants and Test 2 included 1,003,453. Imputation quality from findhap 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 in common with HD genotypes and then imputed back to sequence. Predictions were tested using HD imputed genotypes for 26,970 progeny-tested bulls and 2015 data of 3,983 validation bulls with daughters that were first phenotyped after August 2011. Percentage of correctly imputed variants averaged 97.2% across all chromosomes in Test 1 and 97.0% in Test 2. Prediction reliability improved only 0.6 percentage points in Test 1 when sequence SNPs were added to HD SNPs and was only 0.4 points higher than HD SNPs in Test 2 when sequence SNPs and indels were included. However, selecting the 16,648 candidate SNPs with largest estimated effects and adding those to the 60,671 SNPs used in routine evaluations improved reliabilities by 2.7 percentage points (67.4% vs. 64.7%) on average across traits compared with 35.2% for parent average reliability. Thus, genomic prediction reliabilities can improve when adding selected sequence variants.

Key words: genomic prediction, reliability, sequence variant, whole genome sequencing

Introduction

Accuracy of genomic predictions can be improved by using more variants, including variants pre-selected for effects, or including variants near genes, 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 (QTLs) or preselected variants more closely linked to the QTLs. Nearly 40 million variants have been identified from whole genome sequence (WGS) data for >1,500 bulls, and several strategies show potential for imputing these variants to additional animals and using them in genetic evaluation for economic traits (Brøndum *et al.*, 2014, 2015; Druet *et al.*, 2014; van Binsbergen *et al.*, 2014, 2015; Pérez-Enciso *et al.*, 2015; Calus *et al.*, 2016; MacLeod *et al.*, 2016). Candidate variants can be targeted to specific traits such as genes related to fertility, thereby improving reliability for daughter pregnancy rate by 0.20 percentage points when 39 SNPs were added to the marker set used for genomic prediction (Ortega *et al.*, 2016). Numbers of sequenced animals should continue to increase as researchers examine more families and the costs of generating data continue to decline.

Imputing, selecting, and predicting effects for millions of variants and many thousands of individuals requires efficient computation (VanRaden and O'Connell, 2015). Computational costs proportional to the number of variants multiplied by individuals could exceed the marginal benefits from adding more variants. Variants in or near genes should improve reliability of predictions, and direct use of causal variants is preferred to using linked markers. Strategies to choose variants to include on genotyping arrays of different densities or in routine predictions were developed and compared using simulated data for Holstein bulls. The research reported here examined simulated data first and then actual sequence genotypes from the 1000 Bull Genomes Project (Hayes *et al.*, 2014).

The goals of this study are to 1) compare reliability of prediction from sequence data, array data, combined data, and different variant types and 2) investigate edits, imputation, and computing strategies efficient for even larger genotyped populations.

Methods

The SNP and indel calls (sequence variants) from run 5 of the 1000 Bull Genomes Project (Daetwyler *et al.*, 2014) 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 (version 3; http://aipl.arsusda.gov/software/findhap/). Total numbers of variants identified in run 5 were 38 million SNPs and 1.7 million indels, but many of those variants are monomorphic within the Holstein breed. The indels had an average length of 3 and a maximum length of 86. Imputed sequence genotypes from the 1000 Bulls data were set to missing if none of the three genotype probabilities (AA, AB, or BB) were >0.98 as estimated by Beagle (Browning and Browning, 2007).

The HD genotypes of 2,394 Holsteins mainly from North America, Italy, and the United Kingdom were used to impute genotypes of 590,363 other Holsteins that had genotypes from 50K or lower density chips. The imputed HD genotypes of bulls used in this study were a subset of those animals. The original 777,000 HD markers were reduced to 312,614 by removing highly linked markers and other edits before imputation with findhap (version 3). To verify direction and consistency of allele codes, genotypes called from sequences were matched to corresponding chip markers for 155 Holstein or red Holstein animals that had chip genotypes imputed in the U.S. database and sequences in the 1000 Bull Genomes database.

Variants with a minor allele frequency (MAF) of <0.01, incorrect map locations, excess heterozygotes, or low correlations of sequence and HD genotypes for the same variant were removed. After merging sequence and HD data, Mendelian conflicts between parents and progeny were set to missing for 0.01% of 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 allele frequencies were substituted for those when used in genomic prediction. All HD markers that were also in the sequence data were kept except in cases where the absolute correlation among HD markers was <0.95. This edit removed <1,000 (0.3%) of the HD markers because a similar edit had previously been applied before imputation (VanRaden *et al.*, 2013). A few hundred sequence variants were removed in specific regions already known to be mapped incorrectly in UMD3.1.

Three different variant sets were imputed, testing the use of candidate SNPs (Test 1), candidate SNPs and indels (Test 2), or also including intergenic and intronic variants (Test 3). The initial sequence genotype edits 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 VCF file contains three genotype probabilities from Beagle, and the edit for Tests 1 and 2 simply took any genotype with probability of >0.98. The new edits used a probability of >0.95, and after extracting data, a second edit improved input data quality across individuals by deleting any variant with >5% missing genotypes for low frequency variants (<10% MAF) or >MAF/2 missing for more common variants. A third new edit for Test 3 deleted variants with more than 1.5 of the expected $2p(1-p)$ heterozygotes. Only 3,148,506 variants remained after these edits that were added to improve imputation accuracy for all samples.

Quality and orientation of calls were examined using 179 bulls that had both sequence and HD genotypes. After reversing the orientation of HD markers to match sequence and keeping the sequence instead of the HD genotypes for animals that had both, the two data sets were combined for a total of 27,235 animals. Imputation quality 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 in common with HD markers and then imputed back to sequence. The percentage of imputed genotypes that matched the original genotypes was the simple measure of success.

Genomic predictions were computed using deregressed evaluations from August 2011 for 33 traits and 19,575 bulls. Predictions were tested using later data of 3,983 bulls with daughters that were first phenotyped after August 2011. 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 exonic, 9,422 splice, 35,242 untranslated regions at the beginning and ending of genes, 254,907 within 2 kb upstream, and 74,862 within 1 kb downstream variants for a total of 481,904 candidate variants based on Ensembl gene annotation. Test 2 also included any indels located between 2 kb upstream and 1 kb downstream. Test 3 imputed data were used only for genome-wide association (GWA) because genomic predictions converged too slowly with >3 million variants, and GWA results from actual data will be reported in a separate article.

A subset of variants were selected for potential use in a routine genomic prediction using methods similar to those used previously to select HD markers with largest effects in the national evaluation (Wiggans *et al.*, 2016). The 16,648 sequence variants with the largest effects were selected from the analysis of 762,588 and added to the 60,671 markers used previously. However, 6,584 of those previously used markers were not present in the sequence data and were not included in the final set of 70,735 tested.

Results and Discussion

Sequence Variants

Edits for sequence variants are documented in Table 1. Twenty million of the initial 39 million variants were removed for low MAF, and another 13 million were removed because of high linkage with neighboring 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.

Only 91% of the 60,671 markers currently used in official U.S. evaluations were present in sequence data. Some markers with low MAF might be expected to be missing, but the average MAF of the 9% that were missing and the 91% that matched were both about 0.28 for Holsteins. The missing markers are evenly scattered across the chromosomes and probably do not indicate reference genome misassemblies but are likely due to edits during variant identification (Daetwyler *et al.*, 2014). The individual correlations of HD with sequence genotypes were mostly near +1 or –1, which indicates good quality for the 91% of HD markers present in the sequence data. About half of the genotypes had opposite allele coding compared to the sequence variant calls because variants in sequence data were coded based on differences from a Hereford cow-derived reference genome, whereas a preset Illumina array manifest file was used for array allele coding.

The percentage of variants correctly imputed in Test 1 averaged 97.2% of 762,588 across all chromosomes, with a maximum of 98.5% for BTA20 and BTA22 and minimums of 94.9% for BTA15 and 95.0% for BTA4 (Figure 1). The X chromosome was split into the pseudo-autosomal region (labelled as BTA30) with poor imputation) and the X-specific loci (labelled as BTA31); no Y loci were present. Imputation accuracy was slightly reduced to 97.0% with the 1,003,453 variants including indels in Test 2 and to 96.7% with the 3,148,506 variants including intronic and intergenic variants in Test 3. These percentages are inflated because they include the HD markers that were already present. The low imputation accuracy for chromosome 12 in Test 3 was mainly caused by a gap from 72.4 to 75.2 Mb where no markers were available from the HD

Table 1. Edits applied to actual data in Test 3

Edit category	Millions
Original number of SNPs called	39
Removed for MAF of <0.01	20
Removed for linkage of >0.95	13
Removed for imputation accuracy	3
Remained after edits	3

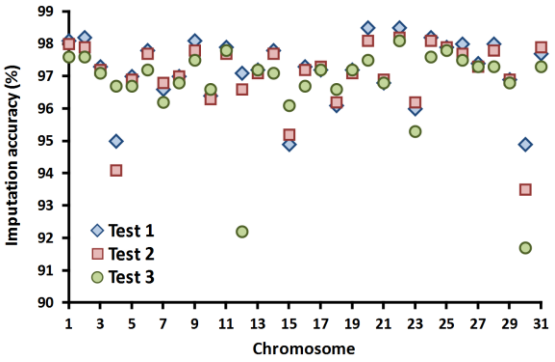

Figure 1. Accuracy by chromosome of imputing sequence variants for 762,588, 1,003,453, and 3,148,506 variants in three tests. Chromosome 30 refers to the pseudo-autosomal region of X, and 31 refers to X-specific loci.

array. Total time required to prepare, edit, and impute the 762,588 variants for 27,235 animals ranged from 1 to 5 hours per chromosome (Table 2) and about 5 days for all 30. Data manipulation steps such as transposing the sequence data and merging with HD data used 1 thread and took more time than the imputation, which used 20 threads and took <1 day.

Reliability of predictions improved only 0.6 percentage points on average using the 762,588 sequence variants and HD data compared with using HD data only (Table 3). Inclusion of indels decreased the advantage over HD to only 0.4 percentage points. Compared with the 60,671 SNPs used currently, reliability improved by about 2.7 percentage points for the final set of 70,735 variants, which included the 60,671 minus the 6,584 not included in the sequence data plus the 16,648 sequence variants selected with largest effects. Reliabilities were 35.2% from parent average, 64.7% from 60,671 SNPs, 67.4% from 70,735 variants, 64.0% from HD SNPs, 64.6% from HD plus genic SNPs, and 64.4% from HD plus genic SNPs and indels. The 60,671 already included the best SNPs selected from HD SNPs (Wiggans *et al.*, 2016), which may explain why 60,671 slightly outperformed HD.

For use with lower density genotyping arrays (5K), the 16,648 sequence variants were further restricted to 4,822. Hand edits were applied to prevent too many candidate SNP from all tagging the same QTL. Figure 2 provides an example for chromosome 5 of the SNPs kept and removed. The same list of 4,822 SNPs was

Table 2. Time (minutes) required with actual sequence data to complete each computational step for longest (BTA1) and shortest (BTA29) chromosomes.

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	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	1	1
Impute with reduced data	17	9
Total time	**296**	**78**

provided to Zoetis (Florham Park, NJ), GeneSeek (Lincoln, NE), and Genetic Visions (Middleton, WI) for potential inclusion on revised arrays. The benefits of adding the sequence SNPs directly to lower density rather than only to medium or higher density arrays are to genotype more young animals quickly and to avoid imputation loss when including sequence SNPs in routine predictions. Re-genotyping or sequencing more reference animals could also help avoid imputation loss in SNP effect estimation for newly discovered variants.

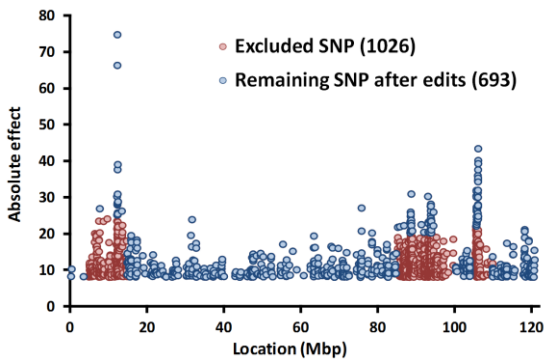

Figure 2. Example variants selected from chromosome 5; only those with larger effects were kept in windows containing the largest effects.

Table 3. Reliability gains (percentage points) over parent average (PA) when adding actual sequence variants to HD or to 60,671 SNPs.

	HD + candidate SNPs			60,671 markers + selected SNPs			PA	
Trait	HD only	HD + 481,904	Difference	60,671 only	60,671 + 16,648	Difference	reliability (%)	HD + indels
Milk	34.1	33.9	–0.2	34.3	35.7	1.4	37.9	33.9
Fat	33.7	34.0	0.3	34.3	35.1	0.8	37.9	33.4
Protein	27.9	27.0	–0.9	27.5	28.2	0.7	37.9	26.7
Fat percentage	49.2	52.7	3.5	52.9	54.8	1.9	37.9	52.4
Protein percentage	42.1	41.6	0.5	41.6	44.3	2.7	37.9	43.0
Productive life	36.1	35.8	–0.3	35.6	38.2	2.6	32.0	36.4
Somatic cell score	35.9	36.1	0.2	35.1	37.0	1.9	34.7	37.1
Daughter pregnancy rate	30.8	30.0	–0.8	29.0	33.0	4.0	31.5	31.2
Cow conception rate	28.7	28.1	–0.6	28.9	31.8	2.9	29.8	28.8
Heifer conception rate	19.0	20.3	1.3	20.5	21.5	1.0	30.0	19.7
Sire calving ease	27.8	27.7	–0.1	24.5	28.5	4.0	29.9	25.2
Daughter calving ease	32.5	30.8	–1.7	31.5	31.4	–0.1	25.3	29.9
Sire stillbirth	7.6	7.3	–0.3	7.6	7.8	0.2	29.0	7.1
Daughter stillbirth	37.4	37.0	–0.4	35.4	38.0	2.6	23.8	35.8
Final score	24.7	25.5	0.8	24.6	27.8	3.2	36.2	25.8
Stature	30.4	32.4	2.0	30.3	34.7	4.3	38.2	32.8
Strength	29.9	31.8	1.9	29.9	34.5	4.6	37.4	31.8
Dairy form	33.8	35.3	1.5	35.0	38.2	3.2	37.4	35.8
Foot angle	17.3	17.6	0.3	17.2	19.6	2.4	36.7	18.2
Rear legs (side view)	21.9	22.7	0.8	22.1	24.1	2.0	37.3	22.0
Body depth	31.0	33.1	2.1	31.2	36.0	4.8	37.6	33.7
Rump angle	32.7	34.0	1.3	32.9	36.1	3.2	37.8	33.5
Rump width	29.2	30.4	1.2	29.1	32.5	3.4	37.1	30.2
Fore udder attachment	35.1	36.4	1.3	35.0	39.0	4.0	37.5	36.1
Rear udder height	24.7	25.7	1.0	24.1	27.3	3.2	37.3	25.8
Udder depth	40.2	42.6	2.4	40.6	44.6	4.0	38.0	42.8
Udder cleft	23.7	24.5	0.8	23.6	25.5	1.9	37.1	24.0
Front teat placement	32.6	33.4	0.8	30.9	35.0	4.1	37.6	32.3
Teat length	29.0	30.3	1.3	28.0	32.7	4.7	37.7	29.9
Rear legs (rear view)	20.7	20.3	–0.4	20.4	22.8	2.4	36.0	20.1
Feet and leg score	16.9	16.5	–0.4	15.9	18.3	2.4	36.4	16.6
Rear teat placement	33.1	33.6	0.5	32.9	35.2	2.3	37.4	32.1
Net merit	23.8	24.3	0.5	23.4	24.7	1.3	34.4	24.4
Average	**28.8**	**29.4**	**0.6**	**29.5**	**32.2**	**2.7**	**35.2**	**29.2**

Comparison with Previous Studies

Previous studies used 5,000 bulls with HD SNPs and 10 million variants from run 3 sequence data (van Binsbergen *et al.*, 2015) or 4 million variants from run 4 (Calus *et al.*, 2016), but sequence predictions in those studies had slightly lower reliability than predictions from HD genotypes only or from BovineSNP50 genotypes only. The HD genotypes in those studies were all observed, but HD genotypes used in our study were mostly imputed. Use of sequence data from run 3 or 4 instead of run 5 could explain their small negative instead of positive gains. The similar results from their studies and ours suggest that errors in the sequence data variants or remaining reference assembly mistakes that altered the order of variant sites in the sequence data could account for the small changes in reliability of prediction.

Our results indicate that adding selected sequence variants can be useful in routine prediction even if analysis of all variants is not more accurate or feasible, which is consistent with previous conclusions for sequence data (Brøndum *et al.*, 2015) or HD data (Saatchi and Garrick, 2014; Wiggans *et al.*, 2016). Brøndum

et al. (2015) added 1,623 sequence variants selected by GWA from multiple breeds to a custom chip and reported average reliability gains of about 2 percentage points. Small improvements (0.2 points) from adding SNPs located in genes associated with fertility were observed by Ortega *et al.* (2016), which is consistent with gains reported in this and earlier studies (VanRaden *et al.*, 2013). Using sequence data and giving extra weight to candidate variants can improve predictions across breeds (Iheshiulor *et al.*, 2016; MacLeod *et al.*, 2016; van den Berg *et al.*, 2016a, 2016b), but advantages of focusing on candidate variants decrease if not all QTLs are in the variant set (Pérez-Enciso *et al.*, 2015). Multi-trait methods can detect QTLs that single-trait methods might miss (Pausch *et al.*, 2016), and even uncorrelated traits can help separate QTLs from markers if many independent traits are controlled by a limited number of QTLs.

Linkage disequilibrium and MAF distributions in the 1000 Bull Genomes sequence data are shown in Figures 3 and 4, respectively. Edits for MAF and for high linkage disequilibrium reduced the 39 million actual variants to 6.3 million (Table 1). Our edits were similar to those of Calus *et al.* (2016), who obtained 4.1 million variants from Holstein data in run 4.

Reliability gains from actual sequence data were higher than previous gains reported from HD data. Larger gains may be possible if the selected SNPs are added to highly accurate arrays and genotyped directly instead of imputed from less accurate sequence data. Accuracies of genotypes from sequence variant calling can vary (Baes *et al.*, 2016), whereas the error rate of Illumina BeadChip arrays is <1% for nearly all SNPs.

Computation

Most computing steps in Table 2 were programmed in Fortran for efficiency, but several steps were in SAS for convenience. The SAS program to merge sequence and HD data took only 6 minutes for the shortest chromosome but 143 minutes for the longest; this program 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 total time. 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, plus additional progress globally for foreign breeders that directly use the new genotyping arrays or that indirectly benefit by selecting breeding stock from the improved U.S. population. Domestic progress is now about $50 per cow annually and would increase to $51 after multiplying by the accuracy ratio of 1.02 which equals the square root of the reliability ratio (67.4/64.7). This higher

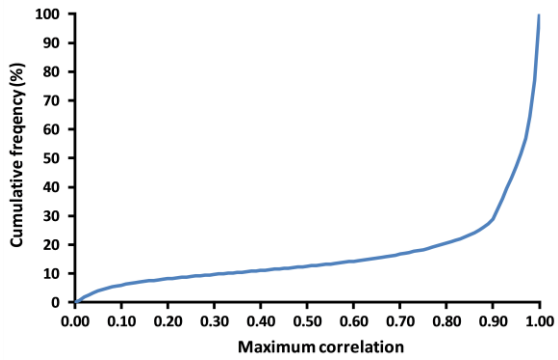

Figure 3. Maximum correlations with neighboring variants in the sequence data.

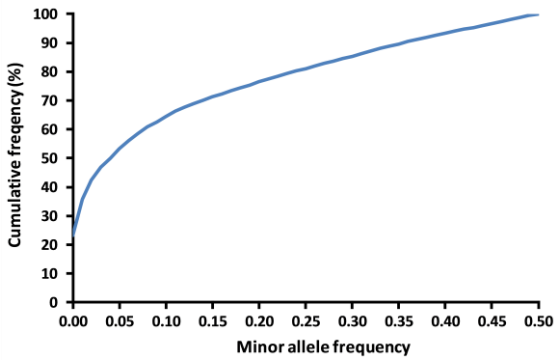

Figure 4. Cumulative distributions for MAF in the sequence data.

accuracy has an annual national value of about $3 million because each year 3.3 million of the 9.2 million U.S dairy cows are replaced, and these annual gains are permanent and will accumulate. The initial cost of generating the U.S. sequence data for the 88 dairy bulls contributed to the 1000 Bull Genomes Project was $132,000 at current reagent costs (~$1,500 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 these tests using imputed data, but most reference animals will still have imputed data. Re-genotyping old animals with the new arrays might be less expensive than more sequencing to improve imputation accuracy.

Conclusions

Variant selection is needed because all of the millions of sequence variants for all animals cannot be imputed and included in routine genomic predictions. Large gains in reliability are possible if the true QTLs can be identified or if bioinformatics methods can choose regions more likely to contain causative variants. Large reference populations are needed in either case because individual QTLs have such small effects. Testing many individual traits gives more power because effects of each QTL may be detectable only for a few traits, but these same QTLs often affect several correlated traits. Assigning more prior variance to the QTLs or to the newly selected variants can improve reliability when estimating effects, but the markers from previous arrays must be retained during imputation because genotypes of previous animals include only the markers 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 >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. Imputation allows many more sequence variants to be tested, selected, and included in routine predictions to increase their reliability. Gains from selecting and including candidate sequence variants were larger than from selecting HD markers.

References

Baes, C.F., Bapst, B., Seefried, F.R., Flury, C., Signer-Hasler, H., Garrick, D.J., Stricker, C. & Gredler, B. 2015. Across-breed imputation with whole genome sequence data in dairy cattle. *Proceedings of Plant & Animal Genome XXIII, San Diego, CA*, Abstract P0281. https://pag.confex.com/pag/xxiii/webprogram/Paper16562.html

Brøndum, R.F., Guldbrandtsen, B., Sahana, G., Lund, M.S. & Su, G. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics 15*, 728. doi:10.1186/1471-2164-15-728

Brøndum, R.F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D. & Lund, M.S. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science 98*, 4107–4116. doi:10.3168/jds.2014-9005

Browning, S.R. & Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics 81*, 1084–1097. doi:10.1086/521987

Calus, M.P.L., Bouwman, A.C., Schrooten, C. & Veerkamp, R.F. 2016. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. *Genetics Selection Evolution 48*, 49. doi:10.1186/ s12711-016-0225-x

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, D., Bouchez, O, Rossignol, M.-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., Van Tassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R. & Hayes, B.J. 2014. Whole-genome sequencing of 234

bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics 46*, 858–865. doi:10.1038/ng.3034

Druet, T., MacLeod, I.M. & Hayes, B.J. 2014. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity 112*, 39–47. doi:10.1038/hdy.2013.13

Hayes, B.J., MacLeod, I.M., Daetwyler, H.D., Bowman, P.J., Chamberlain, A.J., Vander Jagt, C.J., Capitan, A., Pausch, H., Stothard, P. Liao, X., Schrooten, C., Mullaart, E., Fries, R., Guldbrandtsen, B., Lund, M.S., Boichard, D., Veerkamp, R.F., Van Tassell, C.P., Gredler, B., Druet, T., Bagnato, A., Vilkki, J., de Koning, D.-J., Santus, E. & Goddard, M.E. 2014. Genomic prediction from whole genome sequence in livestock: The 1000 Bull Genomes Project. *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, Vancouver, BC, Canada*, Communication 183. https://asas.org/docs/default-source/wcgalp-proceedings-oral/183_paper_10441_manuscript_1644_0.pdf

Iheshiulor, O.O.M., Woolliams, J.A., Yu, X., Wellmann, R. & Meuwissen, T.H.E. 2016. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genetics Selection Evolution 48*, 15. doi:10.1186/s12711-016-0193-1

MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., Chamberlain, A.J., Schrooten, C., Hayes, B.J. & Goddard, M.E. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics 17*, 1–21. doi:10.1186/s12864-016-2443-6

Ortega, M.S., Denicol, A.C., Cole, J.B., Null, D.J. & Hansen, P.J. 2016. Use of single nucleotide polymorphisms in candidate genes associated with daughter pregnancy rate for prediction of genetic merit for reproduction in Holstein cows. *Animal Genetics 47*, 288–297. doi:10.1111/age.12420

Pausch, H., Emmerling, R., Schwarzenbacher, H. & Fries, R. 2016. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. *Genetics Selection Evolution 48*, 14. doi:10.1186/s12711-016-0190-4

Pérez-Enciso, M., Rincón, J.C. & Legarra, A. 2015. Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised. *Genetics Selection Evolution 47*, 43. doi:10.1186/s12711-015-0117-5

Saatchi, M. & Garrick, D.J. 2014. Improving accuracies of genomic predictions by enriching 50K genotypes with markers from 770K genotypes at QTL regions. *Proceedings of the ADSA-ASAS Midwest Meeting, Des Moines, IA*, Abstract 14. https://asas.confex.com/asas/mw14/webprogram/Paper1922.html

van Binsbergen, R., Bink, M.C.A.M., Calus, M.P.L., van Eeuwijk, F.A., Hayes, B.J., Hulsegge, I. & Veerkamp, R.F. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution 46*, 41. doi:10.1186/1297-9686-46-41

van Binsbergen, R., Calus, M.P.L., Bink, M.C.A.M., van Eeuwijk, F.A., Schrooten, C. & Veerkamp, R.F. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution 47*, 71. doi:10.1186/s12711-015-0149-x

van den Berg, I., Boichard, D., Guldbrandtsen, B. & Lund, M.S. 2016a. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: A simulation study. *G3 Genes Genomes Genetics 6*, 2553–2561. doi:10.1534/g3. 116.027730

van den Berg, I., Boichard, D. & Lund, M.S. 2016b. 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. *Journal of Dairy Science 99*, 8932–8945. doi:10.3168/jds.2016-11073

VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., van Kaam, J.B.C.H.M., Valentini, A., Van Doormaal, B.J., Faust, M.A. & Doak, G.A. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science 96*, 668–678. doi:10.3168/jds.2012-5702

VanRaden, P.M. & O'Connell, J.R. 2015. Strategies to choose from millions of imputed sequence variants. *Interbull Bulletin 49*, 10–13. https://journal.interbull.org/index.php/ib/article/view/1367/1435

Wiggans, G.R., Cooper, T.A., VanRaden, P.M., Van Tassell, C.P., Bickhart, D.M. & Sonstegard, T.S. 2016. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. *Journal of Dairy Science 99*, 4504–4511. doi:10.3168/jds.2015-10456