Strategies to Choose from Millions of Imputed Sequence Variants

P. M. VanRaden[1] and J. R. O'Connell[2]

[1] Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA

[2] University of Maryland School of Medicine, Baltimore, MD 21201, USA

Abstract

Millions of sequence variants are known, but subsets are needed for routine genomic predictions or to include on genotyping arrays. Variant selection and imputation strategies were tested using 26 984 simulated reference bulls, of which 1 000 had 30 million sequence variants, 773 had 600 000 markers, 24 863 had 60 000 markers, and 348 had 12 000 markers. Edits for minor allele frequency (MAF) of >0.01, linkage disequilibrium of <0.95 and keeping all 0.5 million variants in or near genes reduced the list to 8.4 million, and those were imputed for all bulls. Strategies were compared to choose variants most significant or with largest estimated variances or effect sizes for five independent traits using single or multiple regression. Reliability of prediction averaged 28.4% from parent average, 77.8% from 60 000, 80.1% from 600 000, 85.0% from 60 000 plus the best 25 000 selected sequence variants or 87.2% using only the 10 000 imputed true quantitative trait loci (QTLs) with no weight on the markers. Genome-wide association (GWA) was faster for selecting variants, but multiple regressions were more reliable. With many genotyped animals and many data sources, computing strategies must efficiently balance costs of imputing, selecting and predicting when millions of variants are available.

Key words: genomic evaluation, whole genome sequence, imputation, variant selection

Introduction

Accuracy of genomic predictions can be improved by using more markers, including markers pre-selected for effects or including variants near genes, within genes, predicted to affect gene function or known to be causal. Nearly 40 million variants have been identified from whole genome sequences for >1 500 bulls, and several strategies to impute and use these show potential (Brøndum *et al.*, 2014, 2015; Druet *et al.*, 2014; Pérez-Enciso *et al.*, 2015; van Binsbergen *et al.*, 2014a, 2014b). Numbers of sequenced animals should continue to increase as researchers examine more families and costs decline.

Imputing, selecting and predicting effects for millions of variants all require efficient computation. Direct use of known QTLs or selecting variants in or near genes can improve reliability of predictions. 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. Their actual sequences from the 1000 Bull Genomes Project (Hayes *et al.*, 2014) were not yet available to us at the time of this simulation study but have since become available.

Methods

Sequence variants were simulated for the 26 984 Holstein bulls in the U.S. reference population in December 2014 using their same pedigree file of 112 905 animals. The 1 000 bulls with the most daughters had 30 million sequence variants, whereas 773 were reduced to 600 000 (600K) markers, 24 863 to 60 000 (60K) markers, and 348 to 12 000 markers to mimic their actual available genotypes. Each simulated chip was an evenly spaced subset of the previous chip. Breeding values were summed for five independent traits from effects of 10 000 (10K) loci, using a heavy-tailed distribution such that the largest effect contributed 3 to 13% of genetic variance, the largest 10 effects contributed 20 to 34%, the largest 100 contributed 57 to 63%, and largest 1 000 contributed 90 to 93%.

The variant list was initially edited to have MAF of >0.01 and to reduce linkage disequilibrium. If any of the 350 variants on either side were correlated by >0.95 absolutely, one variant from each group was retained to represent the group. The 600K markers were all retained to improve imputation, and the 505 210 single nucleotide polymorphisms (SNPs) that were within 2 500 bases of a true QTL were retained to mimic bioinformatic selection using gene positions. After imputing the 8.4 million edited variants for all bulls, GWA chose the most significant 5 000 variants for five independent traits. Simulated phenotypes had reliabilities equal to those of the actual bulls in both variant selection and genomic prediction. The oldest 17 896 bulls were the reference population, and true breeding values of the 9 088 younger bulls were used for validation.

Variants can be selected for highest significance test, largest absolute effect or genetic variance contributed by the locus, which is computed as $2p(1 - p)\text{effect}^2$, where p is allele frequency. Selecting those variants that contribute the most variance has more theoretic appeal and also chooses variants with higher MAF, which could help with imputation accuracy. Using GWA, significance of each variant was tested conditional on neighboring variants already included, and then the tests were combined for each independent trait into an overall significance. The single regression model in GWA included pedigree rather than genomic relationships. Multiple regression requires many iterations to converge, whereas GWA can test many variants without iteration.

Genomic predictions from 60K or 600K markers were compared to predictions with selected markers added using the same nonlinear A method of VanRaden *et al.* (2013). To mimic the selection process used to design the GeneSeek HD version 1 chip (Wiggans *et al.*, 2014), the top 5 000 high-density markers for each of five traits were selected, and the combined set of 23 600 (24K) selected markers after removing duplicates were added to the 60K markers. To mimic selection on net merit, another test selected markers with the highest variance averaged across five traits instead of selecting top markers for each trait and then combining.

Selecting sequence variants should improve accuracy more than selecting only markers, but the markers must be retained during imputation because sequence variants are not available for most animals. A bioinformatic analysis included the 600K markers plus 500 000 (500K) sequence variants near genes for a total of 1.1 million, similar to the actual analysis of Hayes *et al.* (2014). Another analysis tested adding the 10K true QTLs to the 60K markers, and an upper limit on reliability was obtained using only the imputed QTLs 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. Even higher reliability might be obtained by the more costly process of sequencing or re-genotyping all reference animals for the QTLs instead of using their available marker genotypes to impute their sequence variants.

Results and Discussion

Edits for MAF and linkage disequilibrium removed 3.4 million and 18.4 million variants, respectively, which reduced the variant list from 30 million to 8.4 million that included the 600K markers and the 500K bioinformatic variants. For the 26.6 million variants with MAF of >0.01, maximum absolute correlation with a nearby variant averaged 0.96.

Reliability of prediction averaged 28.4% from parent average, 77.8% from 60K markers, 80.1% from 600K markers or 79.2% from the markers selected by GWA from the 600K-marker chip (Table 1). The reliability gain of 2.3 percentage points for 600K vs. 60K markers is larger than reported earlier from either simulated (0.9) or actual (0.4) genomic predictions (VanRaden *et al.*, 2013). The previous results led to a conclusion that simply

Table 1. Reliabilities (%) from parent average (PA), 60K markers, 600K markers or 60K plus 24K markers selected by GWA from the 600K chip for five traits.

Trait	PA	60K	60K+24K	600K
1	24.4	77.9	79.2	80.3
2	31.2	77.9	79.3	80.1
3	32.7	78.3	79.5	80.4
4	23.3	76.6	77.7	78.6
5	30.4	78.3	80.0	81.2
Average	28.4	77.8	79.2	80.1

adding more markers gave small improvements because prior variance for each marker was smaller and additional markers were imputed rather than directly observed.

Adding 24K markers from the 600K that had largest effects from multiple regression gave higher reliability by 2.2 percentage points than the markers selected by GWA when added to the 60K and also 1.3 percentage points higher than using all of the 600K markers (Table 2), which was consistent with previous results from real data (Wiggans *et al.*, 2014). Selecting markers by effect variance was expected to be better than effect size, but effect size gave slightly higher reliability (81.4 vs. 81.2%). The increased MAF should have improved imputation accuracy, but only 19% of the SNPs were different from the two selection strategies. Selecting 23 000 markers using an average of the five traits had only about 50% of markers in common with the other two strategies and gave slightly lower reliability than selecting for each trait and then combining (81.1 vs. 81.2%).

The bioinformatic analysis of 1.1 million sequence variants produced reliability of 86.4%, much higher than the 81.4% best analysis from selecting 600K markers and only about 1 percentage point less than the 87.2% maximum using just the 10K true QTLs. 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.* (2015). Including 1.1 million variants in routine evaluation or on chips is difficult, but 60K markers plus the top 25 000 chosen from the 1.1 million by multiple regression gave reliability of 85.0%. If the 10K true QTLs were added to the 60K but not given extra prior variance, reliability was only 84.5% because too much prior variance was assigned to the markers.

Table 2. Reliabilities (%) for five traits from 60K plus 24K markers selected from the 600K chip by effect size or effect variance or plus 10K QTLs or from only 10K QTLs (no prior variance for markers).

	Select effect by		Add	Only
Trait	Size	Variance	QTLs	QTLs
1	81.6	81.3	84.6	87.2
2	81.4	81.2	84.9	87.7
3	81.3	81.5	85.0	87.8
4	80.2	79.8	82.9	85.9
5	82.5	82.2	85.2	87.5
Average	81.4	81.2	84.5	87.2

Computing resources for each step are shown in Table 3. Genotype simulation required 56 hr with one processor and 210 GB of memory and output a 32-GB file. Calculation of linkage correlations among neighboring sequence variants and pruning those that were highly correlated took 1 hr with 10 processors and 27 GB of memory. Imputation of 8.4 million variants for 26 984 bulls required 38 hr with 20 processors and 13 GB of memory and output a 220-GB file. Selection of variants by GWA required only a half hour with 30 processors and very little memory. Genomic prediction for 1.1 million variants and five traits required 22 hr with five processors and 20 GB of memory.

Conclusions

Variant selection is needed because routine genomic predictions cannot impute and include all of the millions of sequence variants for all animals. Large gains in reliability are possible if the true QTLs can be identified or if advanced bioinformatic tools can identify regions likely to contain the 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 the QTL may be detectable only for a few traits. Assigning more prior variance to the QTLs or newly selected markers can improve reliability when estimating effects, but the markers from previous chips must be retained during imputation.

Table 3. Computer resources to select from 30 million simulated variants for 1000 sequenced and 25 984 genotyped bulls.

	Proc-	Time	Memory	Disk
Step	essors	(hr)	(GB)	(GB)
Simulate 30 million	1	56	210	32
Prune linkage	10	1	27	10
Impute 8 million	20	38	13	220
Select 25 000	30	0.5	<1	<1
Predict 1 million	5	22	20	<1

Computation becomes a limiting factor as reference populations and target populations grow in size. Total computing time was only a few days with 1 000 sequences and 26 984 bulls, but 150,000 reference cows 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. The same methods tested here will be applied to select variants using the actual sequences and U.S. phenotypes.

References

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.

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.

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.

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.A., Veerkamp, R.F., Van Tassell, C.P., Gredler, B., Druet, T., Bagnato, A., Vilkki, J., deKoning, 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,* Communication 183.

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.

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

van Binsbergen, R., Calus, M.P.L., Bink, M.C.A.M., Schrooten, C., van Eeuwijk, F.A. & Veerkamp, R.F. 2014b. Genomic prediction with 12.5 million SNPs for 5503 Holstein Friesian bulls. *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production,* Communication 664.

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. and Doak, G.A. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science 96,* 668–678.

Wiggans, G.R., Cooper, T.A., Null, D.J. and VanRaden, P.M. 2014. Increasing the number of single nucleotide polymorphisms used in genomic evaluations of dairy cattle. *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production,* Communication 301.