Using the ARS-UCD1.2 reference genome in U.S. evaluations

***D.J. Null*[1]*, P.M. VanRaden*[1]*, B.D. Rosen*[1]*, J.R. O'Connell*[2]*, and D.M. Bickhart*[3]**

[1] *U.S. Department of Agriculture, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705-2350, USA*

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

[3] *U.S. Department of Agriculture, Agricultural Research Service, U.S. Dairy Forage Research Center, Madison, WI 53706, USA*

Abstract

The U.S. genomic evaluation began using a new reference map in December 2018 to improve genotype imputation and lethal carrier detection. Additional selected markers and gene tests also were included at the same time to improve reliability of genomic predictions. After edits removed some markers from both the old and new maps that were potentially mismapped, haplotype inheritance was better with the new map for all five breeds tested. The new map and the inclusion of more gene tests detected more carrier animals for 18 of the 24 individual variants tracked, thereby improving the accuracy of imputing carrier status. A further benefit of using the new map for national genetic evaluation was to allow simpler merging with international sequence data from run 7 of the 1000 Bull Genomes Project that also aligned variants to the new map.

Key words: genomic prediction, reference genome, ARS-UCD1.2, imputation, genotype

Introduction

An updated version of the cattle DNA reference genome (ARS-UCD1.2) was developed by researchers from USDA's Agricultural Research Service, the University of California-Davis, and others (Medrano, 2017; Rosen et al., 2018) to replace the University of Maryland version 3 (UMD3) map used since 2009 by researchers around the world (Zimin et al., 2009). Both maps used DNA sequence from inbred Hereford cow Dominette, but the new map included longer reads to improve accuracy over repetitive sections of DNA. Many cattle researchers are not familiar with the process of updating from one map to another because many used only one map from the beginning.

The U.S. genomic evaluations switched from using UMD3 to using ARS-UCD1.2 in December 2018. International researchers in the 1000 Bull Genomes Project (http://www.1000bullgenomes.com/) also switched to the new ARS-UCD1.2 reference genome instead of UMD3 as the common language for tracking variation in the latest release (run 7) in May 2019.

Properties of the new and previous map for imputation were compared at USDA's Animal Genomics and Improvement Laboratory (AGIL; Beltsville, MD). The list of 60,671 (60K) single-nucleotide polymorphisms (SNPs) previously used by the Council on Dairy Cattle Breeding (CDCB; Bowie, MD) from 2014 to 2018 excluded several sections of UMD3 that were mismapped. Null et al. (2018) initially compared the edited UMD3 map with a pre-release version of the new map. More recent testing used the public version ARS-UCD1.2 plus a further edited version obtained by removing some apparently mismapped regions that caused haplotype non-inheritance. Milanesi et al. (2015) reported small differences in imputation accuracy from using different maps, but an updated map may also improve gene annotation, sequence alignment, and other research properties.

Methods

To test imputation, SNPs were converted from their previous map locations to new map locations. Flanking sequences from array manifests were remapped by R.D. Schnabel (University of Missouri, Columbia, MO) to new locations on ARS-UCD1. They are available at https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/. For a few other SNPs, locations were not available from those files and instead were obtained by aligning the probe sequence or a flanking region near the SNP to the ARS-UCD1.2 map. After using both methods, new locations could still not be uniquely determined for 144 of the 60K SNPs, and those were excluded from ARS-UCD1.2 tests. However, >99% of SNPs aligned in forward direction to the same chromosome. Surprisingly, 142 of the SNPs used are now on different chromosomes (Figure 1). Another 53 SNPs were later designated as mismapped and excluded from the edited test of the ARS-UCD1.2 map based on haplotype properties and low correlations with adjacent SNPs. Thus, 197 previously used markers were not used in the final imputation tests. Additional SNPs were detected and removed using high-density (HD) genotypes in a total of 38 small regions (Appendix 1) and about 9 Mbase or 0.3% of the map (Figure 2) in comparison to previous edits for 10 Mbase from the UMD3 map.

Imputation was performed separately for five different breeds using Findhap version 3. Genotypes from 33 different chips were included from 1,748,453 Holsteins, 215,800 Jerseys, 32,724 Brown Swiss, 4,834 Ayrshires, and 3,517 Guernseys. To test sequence alignment, paired-end reads from a Holstein bull were aligned to both maps.

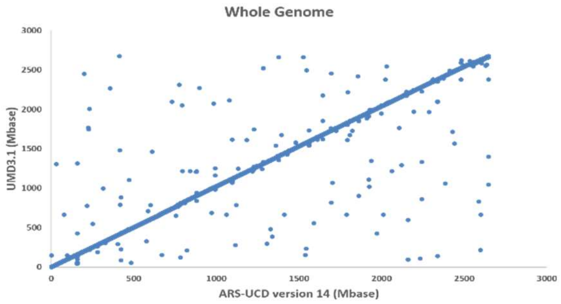

Figure 1. Locations for 142 of 60K SNPs are now on different chromosomes.

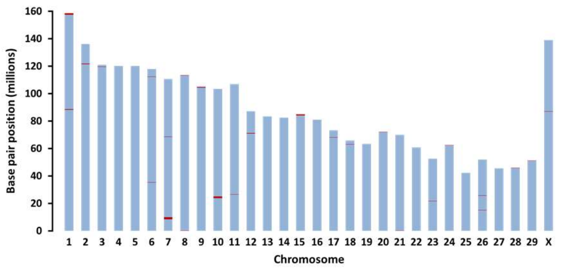

Figure 2. Potentially mismapped sections edited.

Larger Variant Sets

The number of markers used in genomic predictions increased to 79,276 (80K) from the previous 60K used since 2014. The revised list includes more exact gene tests added recently to chips, removes poorer performing markers, adds new variants with larger effects on traits, and changes the marker order based on the new map. Recent chips from Zoetis (Parsippany, NJ), Neogen (Lincoln, NE), and Genetic Visions (Middleton, WI) each include new variants selected by AGIL from sequence or HD chips. Nearly all other countries still use 50,000 or fewer SNPs from the original 50K chip but no gene tests or additional SNPs yet.

Reliability gains from 77,321 SNPs versus the 60K SNP set were estimated in a preliminary study to average 1.4 percentage points across traits for Holsteins when the added SNPs were selected from HD chips including gene tests (Wiggans et al., 2016). Reliability gains were estimated to average 2.7 percentage points when the added SNPs were selected from both sequence and HD data (VanRaden et al., 2017). The final SNP set implemented included a total of 79,294 SNPs and was a combination of these two projects. The final set included about 3,000 instead of 16,000 of the SNPs selected from the sequence data, because only those 3,000 had been added to chips.

Imputation was conducted in two phases for Holstein animals—first by imputing all bulls and their ancestors and then using those haplotypes as priors to impute the remaining 2 million females. The first phase took about 2 days of computing, and the second phase took 1 week with 25 processors and 270 Gbytes of memory (22% of available). The new list of 80K instead of 60K SNPs increased run times for some key programs by about 30%. The new map and the 80K SNP list were then used by CDCB in the December 2018 official evaluations for all traits and breeds and in estimating breed base representation for crossbred cattle.

Results and Discussion

Lower non-inheritance and fewer haplotypes per segment in Tables 1, 2, and 3 indicate that the new map better matches true DNA sequences. Average non-inheritance across all haplotypes was lower for ARS-UCD1.2 than for UMD3 and further decreased after removing 53 apparently mismapped SNPs (Table 1). Maximum non-inheritance rate and maximum number of haplotypes were initially higher for ARS-UCD1.2 than for the edited version of UMD3 for most breeds, but removal of the 53 SNPs reduced maximum number of haplotypes for ARS-UCD1.2 to below that for UMD3 for all breeds (Tables 2 and 3). Many previous problem sections of UMD3 no longer have excess numbers of haplotypes with ARS-UCD1.2, particularly on the X chromosome and the pseudoautosomal region of X, which has been resolved as one contiguous segment (Johnson et al., 2019).

Larger Variant Sets

Large reference populations obtain more benefit from more SNPs because of more phenotypes to estimate each SNP effect. For Holsteins and Jerseys, correlations with previous predicted transmitting abilities are about 0.99, and reliability increases are only about 1 percentage point for yield traits.

One important mutation controlling about 30% of fat yield is now directly included (*DGAT1*; Gautier et al., 2007). Genomic predictions improved the most for the Jersey and Holstein breeds, which have larger reference populations and larger effects of *DGAT1*. More gene tests, quantitative trait loci, selected sequence SNPs, and HD SNPs with larger effects are now included in the SNP list. Now *DGAT1* has larger effects on yield and net merit than any marker in Jerseys and Holsteins and about the same size as the markers near *DGAT1* in Guernseys. As a result, predictions for those breeds changed more than predictions for Ayrshires and Brown Swiss, where *DGAT1* effects were smaller or minor allele frequency was lower.

Numbers of SNPs, inclusion of gene tests, and presence or absence of nearby SNPs with poorer quality can affect carrier status for fertility haplotypes. The new 80K SNP set now contains many more gene tests that were added to recent chips and provided to CDCB, primarily from Neogen. Those tests help impute carrier status for all other animals, but the quality of the gene tests must also be monitored.

Table 1. Average non-inheritance (%) of haplotypes before and after edits.

	Map		
	UMD3	ARS-UCD1.2	
Breed	Edited	Not edited	Edited
SNPs	60,671	60,527	60,474
Removed	0	−144	−197
Ayrshire	1.79	1.52	1.36
Brown Swiss	1.46	1.27	1.16
Guernsey	1.60	1.41	1.31
Holstein	1.56	1.28	1.12
Jersey	4.29	3.93	3.76

Table 2. Maximum non-inheritance (%) of haplotypes before and after edits.

	Map		
	UMD3	ARS-UCD1.2	
Breed	Edited	Not edited	Edited
Ayrshire	16.42	26.58	10.30
Brown Swiss	17.06	21.16	10.38
Guernsey	17.58	16.45	9.45
Holstein	12.88	31.28	12.33
Jersey	24.85	35.47	17.73

Table 3. Maximum number of haplotypes per segment before and after edits.

	Map		
	UMD3	ARS-UCD1.2	
Breed	Edited	Not edited	Edited
Ayrshire	2,030	2,512	1,606
Brown Swiss	11,602	12,846	9,447
Guernsey	1,970	1,657	1,427
Holstein	47,987	81,316	36,690
Jersey	39,628	33,034	29,732

The large change reported for haplotype JH1 in Table 4 was mainly due to gene test edits already announced by CDCB in September 2018. Although the HH5 gene test was intended to be included, it reported many homozygous animals whereas the haplotype had none.

Comparisons of carrier status from the new versus old list in Table 4 reveal that most haplotypes are very stable, but a few more animals switched to being carrier than to being non-carrier. That may result from the gene tests revealing additional families not previously known to be carriers or from better haplotype inheritance with the new map and more rigorous SNP edits. The statistics for Holsteins are from bulls and their ancestors, whereas the status changes more for females with incomplete pedigrees or fewer genotyped ancestors. The statistics for other breeds include all animals. Changes in carrier status occurred more often when carrier frequencies were also high such as for JH1, AH1, and AH2. More changes also occurred for HHR (recessive red) and HH5.

Table 4. Changes in haplotype carrier status with the new map and expanded 80K SNP list.

		Changed to		
Breed haplotype[1]	Same status	Carrier (%)	Non-carrier (%)	Carrier frequency
Holstein				
HH0	99.8	0.17	0.05	3.2
HH1	99.7	0.24	0.05	2.6
HH2	99.3	0.71	0.02	2.6
HH3	99.5	0.46	0.03	4.6
HH4	99.9	0.02	0.01	0.5
HH5	98.2	1.69	0.15	6.2
HH6	…	0.54	0.00	0.5
HHB	99.9	0.03	0.05	0.2
HHC	99.1	0.68	0.14	1.9
HHD	100.0	0.00	0.00	<0.1
HHM	99.9	0.02	0.01	0.1
HHP	99.2	0.57	0.18	3.8
HHR	97.1	1.33	1.53	9.4
HHBR	99.7	0.21	0.08	1.2
HHDR	99.9	0.03	0.00	0.2
HCD	99.1	0.26	0.61	5.9
Jersey				
JH1	98.4	1.27	0.29	18.4
JHP	99.0	0.17	0.79	4.1
Brown Swiss				
BH2	99.4	0.42	0.20	13.3
BHD	98.7	0.92	0.37	3.0
BHM	99.3	0.66	0.06	4.0
BHP	99.6	0.07	0.31	2.5
BHW	99.9	0.07	0.06	1.2
Ayrshire				
AH1	98.2	1.55	0.23	22.2
AH2	99.0	0.69	0.35	21.0

[1] Haplotypes are those reported by Cole et al., 2018.

Conclusions

The ARS-UCD1.2 map showed improved imputation of genotypes. SNPs from several small regions of the final map were removed to control the maximum non-inheritance within individual haplotypes as had been done with the UMD3 map. After these edits, all properties of imputation using ARS-UCD1.2 were better than with the edited version of UMD3. Further research with denser SNP sets and sequence data may reveal additional segments to edit. The new map improves marker locations, and the additional SNPs also improve carrier status detection and genomic predictions.

References

Cole, J.B., VanRaden, P.M., Null, D.J., Hutchison, J.L., Hubbard, S.M. 2018. Haplotype tests for economically important traits of dairy cattle. *AIP Res. Rep. Genomic4(12-18)*.

Gautier, M., Capitan, A., Fritz, S., Eggen, A., Boichard, D., Druet, T. 2007. Characterization of the DGAT1 K232A and variable number of tandem repeat polymorphisms in French dairy cattle. *J. Dairy Sci. 90*, 2980–2988.

Johnson, T., M. Keehan, Harland, C., Lopdell, T., Spelman, R.J., Davis, S.R., Rosen, B.D., Smith, T.P.L., Couldrey, C. 2019. Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. *J. Dairy Sci. 102*, 3254–3258.

Medrano, J.F. 2017. The new bovine reference assembly and its value for genomic research. *Proc. Assoc. Advmt. Anim. Breed. Genet. 22*, 161–166.

Milanesi, M., Vicario, D., Stella, A., Valentini, A., Ajmone-Marsan, P., Biffani, S., Biscarini, F., Jansen, G., Nicolazzi. E.L., 2015. Short communication: Imputation accuracy is robust to cattle reference genome updates. *Anim. Genet. 46,* 69–72.

Null, D.J., VanRaden, P.M., Bickhart, D.M., Cole, J.B., O'Connell, J.R., Rosen, B.D. 2018. Potential benefits from using a new reference map in genomic prediction. *J. Dairy Sci. 101(Suppl. 2)*, 181(abstr. 168).

Rosen, B.D., Bickhart, D.M., Schnabel, R.D., Koren, S., Elsik, C.G., Zimin, A., Dreischer, C., Schultheiss, S., Hall, R., Schroeder, S.G., Van Tassell, C.P., Smith, T.P.L., Medrano, J.F. 2018. Modernizing the bovine reference genome assembly. *Proc. World Congr. Genet. Appl. Livest. Prod., Vol. Mol. Genet. 3*, 802.

VanRaden, P.M., Tooker, M.E., O'Connell, J.R., Cole, J.B., Bickhart, D.M. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol. 49*, 32.

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. *J. Dairy Sci. 99*, 4504–4511.

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., Salzberg, S.L. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol. 10*, R42.

Appendix 1. Potentially mismapped regions of ARS-UCD1.2 that were edited.

chr01:88288254–88589699
chr01:157525745–158534110
chr02:121316059–121751997
chr03:119378630–119559889
chr06:35300301–35344991
chr06:112109597–112324326
chr07:8433502–9956060
chr07:41323000–41434573
chr07:68528026–68619931
chr08:1–234625
chr08:113096592–113273051
chr09:104172188–104634958
chr10:23703142–25059514
chr10:42224617–42224619
chr11:26412888–26626298
chr11:82836074–83006400
chr12:70762931–71197493
chr14:1–146714
chr14: 82285783–82366657
chr15:10979970–11040240
chr15:83996174–85007780
chr17:68078456–68217960
chr18:63017258–63135541
chr20:71850046–71974595
chr21:29986–289503
chr21:58940778–59023509
chr23:21510406–21737841
chr24:62162854–62317253
chr25:13363332–13363334
chr26:14965770–14982852
chr26:15079004–15180612
chr26:15192350–15241107
chr26:25236723–25297707
chr26:25495761–25798796
chr27:116986–116988
chr28:45859111–45940150
chr29:50977673–51098607
chrX:86996907–87035672