J. Dairy Sci. 99:5526–5534
http://dx.doi.org/10.3168/jds.2015-10347

Reducing animal sequencing redundancy by preferentially selecting animals with low-frequency haplotypes

D. M. Bickhart,[1] J. L. Hutchison, D. J. Null, P. M. VanRaden, and J. B. Cole
Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

ABSTRACT

Many studies leverage targeted whole-genome sequencing (WGS) experiments to identify rare and causal variants within populations. As a natural consequence of their experimental design, many of these surveys tend to sequence redundant haplotype segments due to their high frequency in the base population, and the variants discovered within sequencing data are difficult to phase. We propose a new algorithm, called inverse weight selection (IWS), that preferentially selects individuals based on the cumulative presence of rare frequency haplotypes to maximize the efficiency of WGS surveys. To test the efficacy of this method, we used genotype data from 112,113 registered Holstein bulls derived from the US national dairy database. We demonstrate that IWS is at least 6.8% more efficient than previously published methods in selecting the least number of individuals required to sequence all haplotype segments ≥4% frequency in the US Holstein population. We also suggest that future surveys focus on sequencing homozygous haplotype segments as a first pass to achieve a 50% reduction in cost with an added benefit of phasing variant calls efficiently. Together, this new selection algorithm and experimental design suggestion significantly reduce the overall cost of variant discovery through WGS experiments, making surveys for causal variants influencing disease and production even more efficient.
Key words: whole-genome sequencing, redundancy, haplotype, inverse weight selection

INTRODUCTION

Whole-genome sequencing data (**WGS**) is a tool that will be increasingly leveraged for the genomic selection of dairy cattle traits in the near future. Whole-genome sequencing experiments in cattle range from efforts to increase imputation accuracy (Daetwyler et al., 2014), develop genotyping by sequencing (**GBS**) strategies (Elshire et al., 2011), and identify causal variants affecting disease or productive traits (Sonstegard et al., 2013). In all cases, these experiments stand to benefit from improved methods for sample selection to decrease the cost of sequencing and gold standard variant data sets to validate newly discovered variant calls. Although much discussion has been given to such sample selection strategies in the past (Druet et al., 2014; Yu et al., 2014), these methods may not account for biases inherent in individual sequencing platforms themselves, which may affect subsequent analysis and attribution of rare variant calls.

As increasing focus is given to GBS technologies for efficiently genotyping large herds of cattle (Elshire et al., 2011), developing the means to correct for biases inherent in short-read sequencing platforms is of paramount importance. This is particularly the case in scenarios where researchers are considering the use of low-coverage sequence data to reduce costs associated with genotyping animals (Yu and Sun, 2013). To date, the cattle research community lacks the extensively validated pool of variants available to human subject researchers (Abecasis et al., 2012) that have enabled the production of finely tuned WGS variant calling software. Creation of such a resource would allow for the use of a run- and platform-specific filtration method that uses a mixture of models to distinguish between sequencing artifacts and true positive variants (McKenna et al., 2010). Additionally, lists of validated, pre-existing variant sites can be used to increase the rate of alignment of variable sequence reads, thereby hastening the computation of animal genotypes.

The phasing of WGS variant calls represents a substantial challenge, particularly when the target variant calls are of low frequency within the population. The average spacing of SNP markers on common cattle genotyping chips (49.4 kbp for the BovineSNP50 v2; 3.43 kbp for the Illumina BovineHD; Illumina Inc., San Diego, CA) is far too large to allow for sequence read-based phasing (read lengths of 100 to 200 bp) of variants with existing genotype markers (McKenna et al., 2010). To compensate for this deficiency, researchers

Received September 2, 2015.
Accepted November 23, 2015.
[1] Corresponding author: derek.bickhart@ars.usda.gov

may sequence the parents of the sample used in a WGS experiment to guarantee the phasing of heterozygous variant sites by tracing parental origin of haplotypes. This increases the cost of sequencing 3-fold when rare variant sites—such as those predicted to be within haplotypes affecting fertility (VanRaden et al., 2011)—must be properly phased in a target sample.

In this study, we investigated the use of several methods of selecting animals for sequencing that will produce (1) a set of highly validated variants ranging from high to low frequency in the entire registered US dairy herd, (2) phasing of rare variants to improve whole genome imputation, and (3) the most efficient sequencing strategy in terms of sequencing costs. We propose a new strategy for sequencing sample selection, compare it to a previously published method (Druet et al., 2014), and show that substantial gains in efficiency can be made by prioritizing the sequencing of rare haplotypes.

MATERIALS AND METHODS

Data Accession

Individual haplotypes for all 112,113 registered Holstein bulls were retrieved from the Council of Dairy Cattle Breeding's (**CDCB**; Reynoldsburg, OH) US national database. Haplotypes were defined as follows: first, all Holstein animal genotypes were imputed to 60,671 common SNP biallelic markers from several different genotyping arrays using FindHap version 3 (http://aipl.arsusda.gov/software/findhap/) with settings that included 4 iterations and 3 haplotype widths. Use of different size haplotype intermediates in the FindHap algorithm helps improve imputation from lower density chips, but only haplotypes from the short 100-SNP segments were used in choosing bulls. The segments are nonoverlapping, and the list of individual haplotypes is identified from the combinatorial pattern of imputed SNP marker states. Given that the final haplotypes were based on a fixed count of SNP markers, they varied in terms of base pair lengths (average: 4,258,451 bp; standard deviation: 1,484,696 bp). Each haplotype's frequency was estimated from the entire population of Holstein animals in the national database. Due to logarithmic increases in haplotype counts at lower frequency values, we excluded all haplotypes below a 4% frequency threshold in the population.

Sequencing Strategy and Sample Filtration

In all subsequent cases, algorithms were tasked with identifying the minimum number of samples needed to sequence all haplotypes above a 4% frequency threshold. At this frequency threshold, 3,680 haplotypes were present in the national database that were considered for sequencing. Only haplotypes that were present in a homozygous state in target individuals were considered to remove the need to phase variants. To simulate constraints on sample availability, we restricted our search to bulls contained within the Cooperative Dairy DNA Repository's (**CDDR**; Columbia, MO) database that had greater than 3 semen straws in stock.

Random Selection of Samples

As a point of comparison for subsequent algorithms, we selected samples at random (**RAND**) from the filtered list of bulls meeting our CDDR presence requirements. Random selection was implemented using the "PROC SURVEY SELECT" command within SAS version 9.4 (SAS Institute Inc., Cary, NC), and animals were selected until all haplotypes above the 4% frequency threshold were covered by at least one individual. To account for the variability of random selection, we performed 10 replicates of this method and averaged the results.

Maximizing Haplotype Coverage from the Population

We generated 2 implementations of a method proposed by Druet et al. (2014) that maximized the haplotype coverage from the population (termed "AHAP" in the manuscript). To account for the differing objectives of this survey, we made slight modifications to the algorithm to account for different study goals (see Figure 1). The first, labeled AHAP1, is the base implementation of the method to maximize haplotype coverage that was proposed (Druet et al., 2014). We use the same terms as Druet et al. (2014) to avoid confusion. In short, the frequency of ancestral haplotypes is calculated using PHASEBook (Druet and Georges, 2010), and then a score is generated for every individual by summing the frequency of the haplotype at every SNP locus. This score represents the haplotype contribution of each animal and can therefore be used to prioritize individuals for sequencing. We note 2 significant departures from the original implementation of AHAP1: (1) the probability of sequencing depth of coverage for the haplotype [approximated as $(1 - 0.5^{nki})$, where n was the sequencing depth of coverage, k was the frequency of the ancestral haplotype, and i was the SNP marker site] was removed as our survey assumed uniform sequence coverage across samples, and (2) FindHap was used to identify haplotype segments that were used in lieu of PHASEBook (Druet and Georges, 2010) haplotypes. Both AHAP methods used the following equation to generate scores for the efficient selection of animals for sequencing:

$$\text{Sample weight} = \sum_{i=1}^{NHAP} f_i \qquad \text{if } i = \text{homozygous.}$$

In the above equation, f_i is the frequency of the haplotype in the entire national database, as determined by FindHap, and $NHAP$ is the total number of haplotypes. Only haplotypes that are homozygous within the sample were counted toward the sample's weight for selection. After calculating the weight of all samples in the database, sample sequencing priority was assigned by sorting the samples in descending order and selecting the least number of samples that would cover all haplotypes above the 4% frequency threshold.

The AHAP2 implementation differed from AHAP1 by being an iterative approach rather than a strict ranking algorithm. After sample weights were determined, the samples were sorted in descending order based on their cumulative weights. The sample with the highest weight was selected for sequencing and all homozygous haplotypes that the sample contained were removed from consideration from all samples in the next iteration. Sample weights were then recalculated and sorted in descending order. The iteration continued until all homozygous haplotypes above 4% frequency were represented in the selected samples.

Inverse Weight Selection

To preferentially select samples that carried rare frequency haplotypes, we developed an algorithm—inverse weight selection (**IWS**)—that uses an inverted parabolic function to calculate sample sequencing value (weight). Sample weight was determined by the following equation:

$$\text{Sample weight} = \sum_{i=1}^{NHAP} f_i^2 - 2f_i + 1 \quad \text{if } i = \text{homozygous,}$$

where f_i is the frequency of haplotype i in the national database. As f_i approaches zero, the haplotype's score approaches 1. Haplotypes that are more frequent in the database give increasingly smaller cumulative weight to the sample. After cumulative sample weights were calculated, the samples were sorted in descending order and the sample with the highest cumulative weight was selected for sequencing. As in our implementation of AHAP2, all homozygous haplotypes of the sample with the highest weight were excluded from further consideration, and sample weights for all remaining samples in the database were recalculated. The next highest weight sample was selected, and the iteration continued until all homozygous haplotypes at or above 4% frequency were represented.

RESULTS AND DISCUSSIONS

Sample Selection for Sequencing

With the goal of selecting the least number of animals with haplotypes above a specific frequency threshold, we tested several algorithms designed to prioritize individuals for sequencing based on prior SNP genotyping information. In the genomics era of animal breeding, livestock animals have been extensively genotyped using high- and low-density SNP genotyping platforms, making realistic test cases for the use of such algorithms possible. As such, we used genotypes derived from the US national dairy database and haplotypes calculated by the national US dairy genomics evaluation (data provided by the CDCB) as our input data for each selection algorithm. The total number of predicted haplotypes in the national database was 110,588; however, we selected only those haplotypes that were at or above 4% frequency (3,680) to provide a reasonable target for a sequencing project that would be initiated by a funded research group. When considering other haplotype cutoff thresholds lower than 4% frequency, the number of haplotypes that needed to be sequenced increased dramatically. For example, there were 5,096 haplotypes at or above a frequency of 3% (a 40% increase; data not shown). The constant attrition of recombination and de novo mutation likely generated many of the extremely rare frequency haplotypes observed in the database, so a sequencing project that targets all observed haplotypes is likely to be prohibitively expensive and never-ending. Based on the sum total frequency of all haplotypes in the national database, we estimate that the sequencing of all haplotypes ≥4% frequency accounts for 58.7% of the currently observed DNA within the Holstein population in the United States.

Methods for selection varied with respect to their weighting of haplotype frequency in the population, with both AHAP1 and AHAP2 methods selecting high-frequency haplotypes first and IWS preferring lower-frequency haplotypes. A RAND method was also tested to compare the efficiency of each algorithm against an uninformed selection of animals to sequence. To test the utility of these methods, we tested the level of efficiency of each algorithm in selecting animals that represented all considered haplotypes. In the full test, IWS was superior to the other algorithms, as its results selected the fewest individuals (n = 250) that contained all homozygous haplotypes above a 4% frequency in the database. The modified AHAP2 method was the next most efficient method (n = 267), having selected 17 more animals for sequencing over the IWS data set (6.8% more animals than IWS). Surprisingly, the original AHAP1 algorithm was the least efficient method, with results of 5,325 individuals before accounting for all target haplotypes (21.3-fold more than IWS). This significant increase is primarily due to the noniterative nature of AHAP1, which does not rescore animals

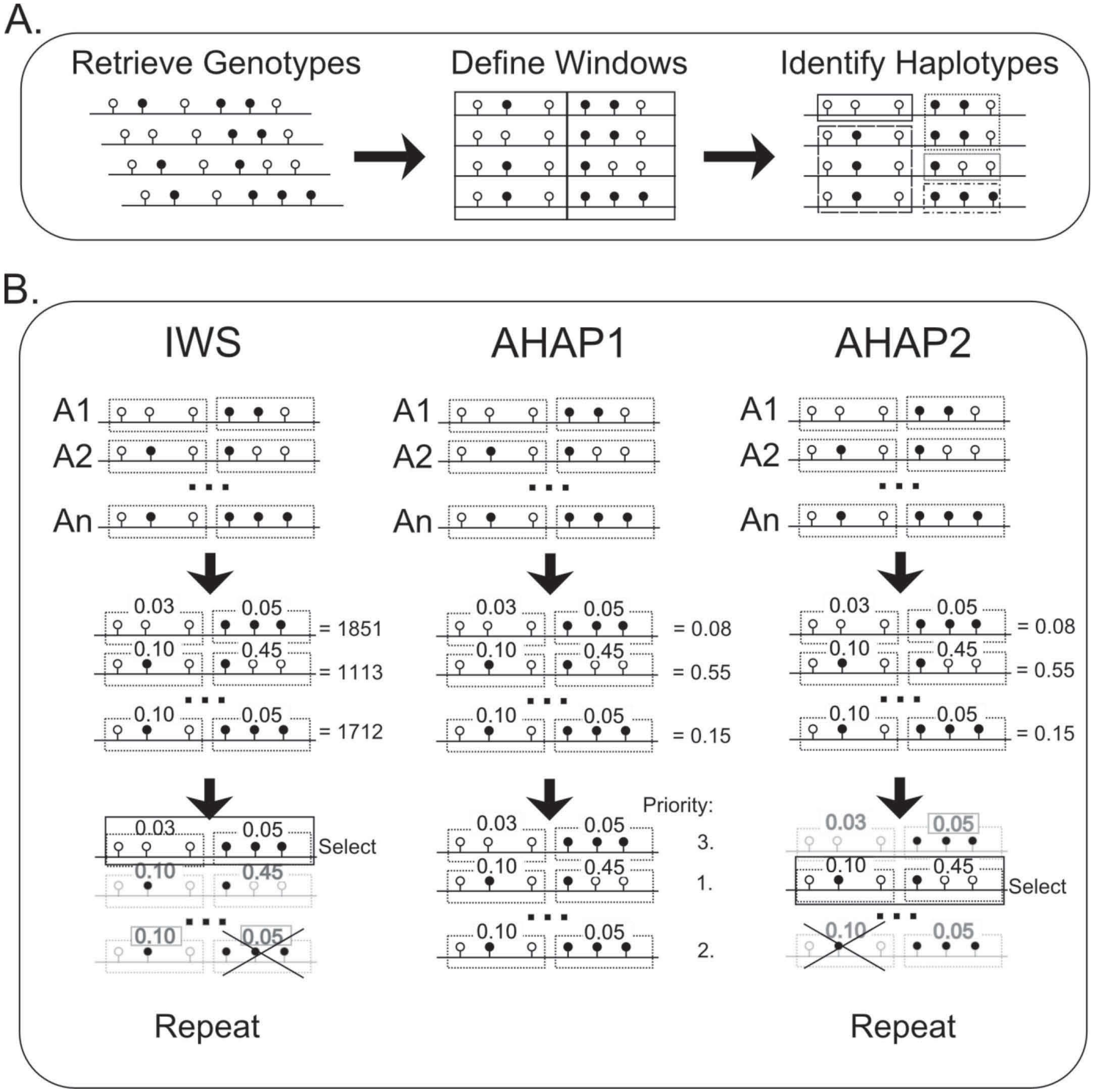

Figure 1. A schematic demonstrating each algorithm, with IWS = inverse weight selection, AHAP1 = maximizing haplotypes coverage from the population, and AHAP2 = modified version of AHAP1 with recursion. The preliminary data accession (A) was similar for each subsequent algorithm, and it resulted in the identification of haplotype segments within the population of sampled individuals (A1, A2,…, An). Each individual algorithm (B) sought to identify the least number of individuals that contained all haplotypes identified in the data accession step.

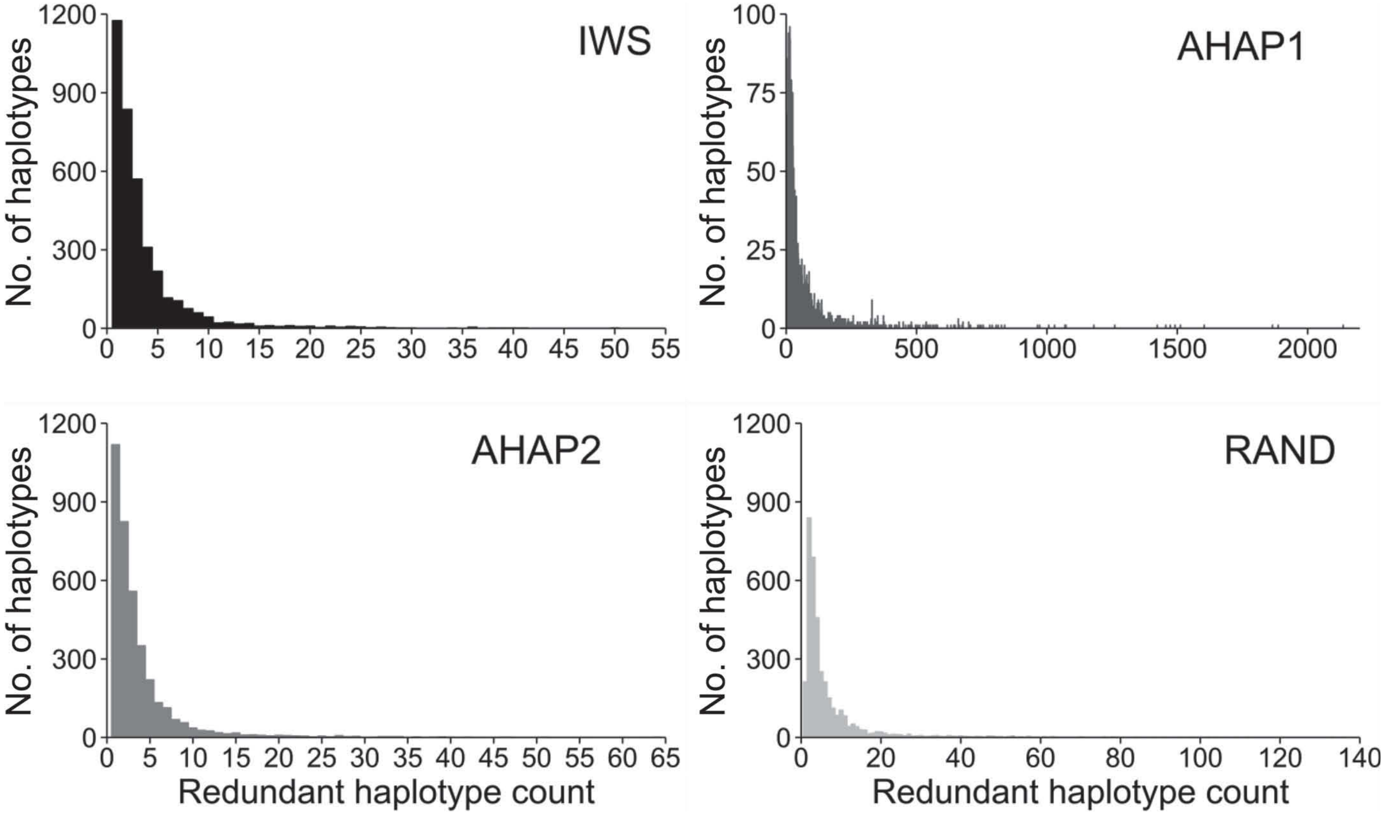

Figure 2. A count of the number of times each haplotype (total haplotype count = 3, 680) is sequenced in passing by 4 algorithms (IWS = inverse weight selection, RAND = random animal selection, averaged over 10 replicates, AHAP1 = maximizing haplotypes coverage from the population, and AHAP2 = modified version of AHAP1 with recursion). The x-axis shows the number of times a haplotype is sequenced, and the y-axis is the count of haplotypes that were sequenced X times.

based on previously selected haplotypes. Adding this feature (as implemented in AHAP2) resulted in a 20-fold improvement in the efficiency of the algorithm with respect to its original implementation. When we compared methods with respect to the number of times that they sequenced the same haplotype segment, we found that the highest count of unique haplotypes per animal was identified using IWS (1,150 haplotypes) out of all other methods (see Figure 2). Haplotypes selected using AHAP1 had higher redundancy than all other methods (average: 661, maximum: 2,137) and was notably worse than RAND.

We also compared the results of our selection algorithms against an approach that preferentially selects highly influential bulls for sequencing. This is a heuristic alternative to a method for selecting influential ancestors (Goddard and Hayes, 2009), insofar as it selects bulls with the most direct descendants. We first selected all bulls in the national database that had more than 5,000 milk-recorded daughters and then sorted the individuals based on their number of daughters. To compare the efficiency of this method against the other algorithms, we treated this data set similarly to our implementation of AHAP1, and selected the minimum number of sequential animals it would take to account for all targeted haplotypes. This approximation of the influential ancestor method required 447 bulls with >5,000 daughters with milk records to account for all haplotypes >4% frequency.

Incidental Rare Haplotype Selection

One fringe benefit of targeted sample selection for WGS may be the incidental sequencing of rare haplotypes (<4% frequency) that were not originally considered in the initial selection calculation. When we accounted for homozygous haplotypes below the initial frequency threshold in the selected animals from each algorithm, we found a substantial number of haplotypes that were sampled in passing (Table 1). The results of AHAP1 incidentally accounted for the largest number of rare frequency, homozygous haplotypes (6,362) whereas AHAP2 and IWS results contained the fewest (3,367 and 3,400, respectively). This relative benefit of AHAP1 is most likely due to the higher number of sampled animals that the algorithm selected rather than any improvement in efficiency. When rare haplotype counts are normalized by dividing the total number of animals selected by each algorithm, IWS (13.6 rare haplotypes per animal sequenced) results showed a nearly 2-fold greater efficiency in incidental selection over RAND (average of 7.7 rare haplotypes per animal sequenced) and a 10% improvement over AHAP2 (average of 12.6 rare haplotypes). When all haplotypes (heterozygous and homozygous) were accounted for, the same trend was observed, with AHAP1 (78.37 rare haplotypes per animal sequenced) results having a lower proportion of sequenced haplotypes than IWS, AHAP2, and RAND (238.21, 232.09, and 166.36, respectively). When incidental haplotypes were considered in estimates of the currently observed DNA present in the Holstein breed, we found that combined IWS and AHAP2 animal haplotypes represented 69.6 and 69.3% of all observed Holstein DNA, respectively. Although several cattle sequencing studies have purportedly accounted for a similar proportion of cattle breed DNA after sequencing a fraction of the number of samples (Jansen et al., 2013; Baes et al., 2014; Daetwyler et al., 2014), the DNA proportion estimates from these studies were made using pedigree-based relationships and do not account for cryptic genetic diversity in these breeds that may arise from common ancestors not present in the pedigree. Therefore, it is likely that the use of these pedigree-based calculations overestimate the true proportion of genetic diversity captured within the sequenced founder animals, and that our estimates using haplotypes derived from dense marker genotypes are more accurate (VanRaden, 2008).

Sequencing Cost Estimates

High sample sequencing costs are a realistic limit to the size of sequencing studies that laboratories can afford to perform. Using a per-megabase cost estimate of $0.05 from recent National Human Genome Research Institute (NHGRI) surveys, the cost to generate 1× coverage cattle WGS data from a sample is approximately $140 (http://www.genome.gov/sequencing-costs/; 2,800 Mb genome size × $0.05/Mb). Because the accuracy of SNP calling is greatly improved by increasing the depth of coverage over a given region of the genome (Bentley et al., 2008; Liu et al., 2012; O'Rawe et al., 2013; Yu and Sun, 2013), realistic experiments would expect to sequence individual, unrelated samples with several-fold coverage to ensure accurate results. Original estimates of required read depth for accurately calling SNP suggested that 15× coverage was needed for homozygous variants and 30× coverage for heterozygous variants (Bentley et al., 2008), suggesting that 2-fold higher coverage is needed to accurately predict heterozygous variants. By focusing on homozygous haplotypes and the predicted homozygous variants within them, sequencing costs can be effectively halved ($2,100 per sample at 15× vs. $4,200 per sample at 30×) by using a lower depth of coverage. Additionally, targeting predicted homozygous haplotypes provides the benefit of removing the need to phase variant calls. Because improvements in SNP calling algorithms have reduced the requirement for higher depth of coverage (Li et al., 2009; McKenna et al., 2010; Yu and Sun, 2013), targeting 6× coverage per sample (~$840 per sample) would further lower costs. Even with the 2 aforementioned cost savings strategies, the most efficient sample selection strategy, IWS, would still require a budget of $210,000 to sequence all haplotypes above 4% frequency in the national database. This represents a savings of $14,280, $342,720, and $4,263,000 over the AHAP2, RAND, and AHAP1 selection methods (Table 2), but still represents a sizeable investment of capital in a single project.

Given that current costs of WGS data are still prohibitively expensive when sampling hundreds of individuals at a time, we estimated the efficiency of each method at different fixed levels of samples sequenced. We predict that many laboratories may only be able to sequence a set number of individuals within a given budget, so this approach gives a fair estimate of the nonredundant genomic information that can be retrieved at different stages of ordered sampling. We selected 4 fixed sample levels (20, 50, 100, and 200

Table 1. Selection performance of different algorithms to account for all haplotypes in the data set

Algorithm[1]	Animals for haplotypes ≥4%	Incidental haplotypes (<4% frequency)	Incidental haplotypes per animal sequenced
IWS	250	3,400	13.6
RAND	658	5,094	7.7
AHAP1	5,325	6,362	1.2
AHAP2	267	3,367	12.6

[1] IWS = inverse weight selection; RAND = random animal selection, averaged over 10 replicates; AHAP1 = maximizing haplotypes coverage from the population; and AHAP2 = modified version of AHAP1 with recursion.

Table 2. Cost efficiency of selection algorithms when sequencing all haplotypes ≥4% frequency

Algorithm[1]	No. of animals	Cost to complete[2]	Average cost per haplotype[3]
IWS	250	$210,000.00	$57.07
RAND	658	$552,720.00	$150.20
AHAP1	5,325	$4,473,000.00	$1,215.00
AHAP2	267	$224,280.00	$60.95

[1] IWS = inverse weight selection; RAND = random animal selection, averaged over 10 replicates; AHAP1 = maximizing haplotypes coverage from the population; and AHAP2 = modified version of AHAP1 with recursion.

[2] The cost to sequence all animals selected by the algorithm, assuming $840 per animal sequenced.

[3] The cost ratio to sequence 3,680 unique haplotypes by sequencing the selected animals in the algorithm.

individuals sequenced, respectively), assumed that each sample level selected individuals in the order of their priority, and counted the number of unique homozygous haplotypes accounted for in each data set at each stage of completion (see Figure 3). In all 4 sample levels, IWS and AHAP2 were superior to the RAND and AHAP1 methods of selection. At the 50-sample threshold (equivalent to 20% of the total IWS animals that account for all haplotypes), IWS-selected animals accounted for 63.4% of the 3,680 haplotypes (Table 3). At the same threshold (50 samples), AHAP2, RAND, and AHAP1 accounted for 61.1, 27.6, and 43.5% of all considered haplotypes. Again, the noniterative nature of AHAP1 hurt the efficiency of the algorithm, though it still accounted for 2,613 haplotypes (71.0% of total) at the 200-sample threshold.

Putative Effect on Imputation

Identifying a high-quality, high-resolution set of variants within cattle sequencing data is likely to improve the accuracy of genotype imputation; however, care must be taken to realize the limitations of this data set with respect to this potential application. The human "1000 Genomes" phase 1 data set was used in the imputation of the Wellcome Trust phase 1 genotype data,

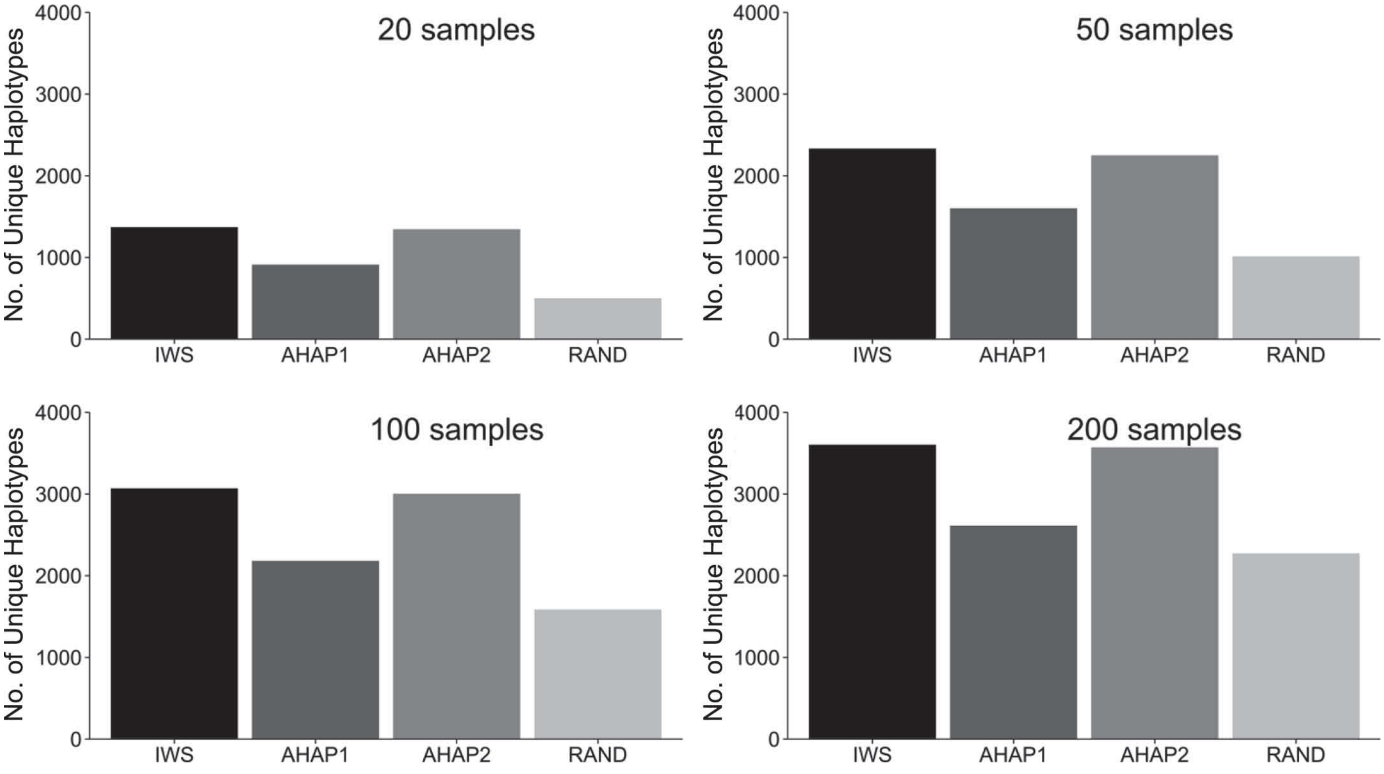

Figure 3. For each algorithm (IWS = inverse weight selection, RAND = random animal selection, averaged over 10 replicates, AHAP1 = maximizing haplotypes coverage from the population, and AHAP2 = modified version of AHAP1 with recursion), the number of haplotypes sequenced at a set stage of sample completion is listed above each bar.

Table 3. Efficiency of tiered completion of sample selection[1]

Samples sequenced	IWS		RAND		AHAP1		AHAP2	
	Hap[2]	H%[3]	Hap	H%	Hap	H%	Hap	H%
20	1,371	37.26	501	13.61	912	24.78	1,345	36.55
50	2,333	63.40	1,014	27.55	1,602	43.53	2,250	61.14
100	3,069	83.40	1,586	43.10	2,182	59.29	3,003	81.60
200	3,604	97.93	2,273	61.77	2,613	71.01	3,571	97.04

[1] IWS = inverse weight selection; RAND = random animal selection, averaged over 10 replicates; AHAP1 = maximizing haplotypes coverage from the population; and AHAP2 = modified version of AHAP1 with recursion.

[2] The number of haplotypes ≥4% frequency represented in the top scoring animals selected at this stage of completion.

[3] The percentage of all considered haplotypes (3,680) that would be accounted for at this tier of completion.

and the resulting association study identified 2 disease-causing variants that were initially overlooked but were confirmed in a follow-up study (Huang et al., 2012). Use of the initial 1000 bulls data from 234 sequenced bulls revealed a high degree of accuracy of imputed calls for SNP with a minor allele frequency >0.1; however, imputation accuracy rapidly decreased for rarer variant sites (Daetwyler et al., 2014). Given that the target haplotype frequency for the current study was 4%, we would expect high accuracy of imputation for variants within the haplotypes at or above that frequency level. The accuracy of imputation would likely decrease for variants lower than 4% frequency, similar to results presented using the 1000 bulls data set. Given that the results from IWS and our implementation of AHAP2 provide scores that represent the novelty obtained by sequencing individual animals, it would be feasible to target lower frequencies of haplotypes by sequencing additional animals. As sequencing costs continue to decline, sequencing additional animals that contain less-frequent haplotypes may become an effective strategy.

CONCLUSIONS

To maximize the utility of WGS for genomic selection, we suggest several strategies for groups interested in using WGS for variant detection that would minimize costs associated with the technology. We have demonstrated that the use of an inverse weight function that prioritizes lower frequency haplotype segments is the most efficient algorithm for selecting a nonredundant set of animals for sequencing. We also suggest that researchers selectively sequence haplotypes that are predicted to be homozygous from SNP genotype data, thereby reducing costs associated with higher depth of coverage sequencing and the phasing of variant calls. These 2 methods correspond to 6.8 and 50% reductions in cost for sequencing projects dedicated to novel variant discovery, respectively. Additionally, we demonstrate that our inverse weight selection algorithm prioritizes animals with higher unique genetic information, which provides high value if only a small portion of the whole population can be sequenced. When budget limitations preclude the ability of researchers to sequence representative animals from an entire population, the use of cost-effective prioritization algorithms and sequencing strategies is the best method for obtaining novel genetic information.

ACKNOWLEDGMENTS

The Council on Dairy Cattle Breeding (Reynoldsburg, OH) and the Cooperative Dairy DNA Repository (Columbia, MO) are acknowledged for providing data used in this study. The authors were supported by appropriated project 1265-31000-096-00, “Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information,” of the Agricultural Research Service of the United States Department of Agriculture (Washington, DC). D. M. B. was also supported by the Agricultural Food Research Initiative (AFRI) competitive grant number: 2015-67015-22970 from the USDA National Institute of Food and Agriculture (NIFA, Washington, DC) Animal Health Program. 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.

REFERENCES

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. http://dx.doi.org/10.1038/nature11632.

Baes, C. F., M. A. Dolezal, J. E. Koltes, B. Bapst, E. Fritz-Waters, S. Jansen, C. Flury, H. Signer-Hasler, C. Stricker, R. Fernando, R. Fries, J. Moll, D. J. Garrick, J. M. Reecy, and B. Gredler. 2014. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. BMC Genomics 15:948 http://dx.doi.org/10.1186/1471-2164-15-948.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, and C. G. Brown. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. http://dx.doi.org/10.1038/nature07517.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. http://dx.doi.org/10.1038/ng.3034.

Druet, T., and M. Georges. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798. http://dx.doi.org/10.1534/genetics.109.108431.

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

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379 http://dx.doi.org/10.1371/journal.pone.0019379.

Goddard, M. E., and B. J. Hayes. 2009. Genomic selection based on dense genotypes inferred from sparse genotypes. Proc. Assoc. Advmt. Anim. Breed. Genet. 18:26–29.

Huang, J., D. Ellinghaus, A. Franke, B. Howie, and Y. Li. 2012. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur. J. Hum. Genet. 20:801–805. http://dx.doi.org/10.1038/ejhg.2012.3.

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pagès, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, and R. Fries. 2013. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics 14:446 http://dx.doi.org/10.1186/1471-2164-14-446.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin., and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.

Liu, Q., Y. Guo, J. Li, J. Long, B. Zhang, and Y. Shyr. 2012. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics 13:S8 http://dx.doi.org/10.1186/1471-2164-13-S8-S8.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. http://dx.doi.org/10.1101/gr.107524.110.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon. 2013. Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med. 5:28 http://dx.doi.org/10.1186/gm432.

Sonstegard, T. S., J. B. Cole, P. M. VanRaden, C. P. Van Tassell, D. J. Null, S. G. Schroeder, D. Bickhart, and M. C. McClure. 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. PLoS ONE 8:e54872 http://dx.doi.org/10.1371/journal.pone.0054872.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. http://dx.doi.org/10.3168/jds.2007-0980.

VanRaden, P. M., K. M. Olson, D. J. Null, and J. L. Hutchison. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J. Dairy Sci. 94:6153–6161. http://dx.doi.org/10.3168/jds.2011-4624.

Yu, X., and S. Sun. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics 14:274 http://dx.doi.org/10.1186/1471-2105-14-274.

Yu, X., J. A. Woolliams, and T. H. E. Meuwissen. 2014. Prioritizing animals for dense genotyping in order to impute missing genotypes of sparsely genotyped animals. Genet. Sel. Evol. 46:46 http://dx.doi.org/10.1186/1297-9686-46-46.