Fast Imputation Using Medium- or Low-Coverage Sequence Data

P. M. VanRaden[1] and C. Sun[2]
[1]Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, Maryland,
[2]National Association of Animal Breeders, Columbia, Missouri, USA

ABSTRACT: Direct imputation from raw sequence reads can be more accurate than calling genotypes first and then imputing, especially if read depth is low or error rates high. An efficient strategy chooses the 2 haplotypes most likely to form the genotype and updates the posterior allele probabilities from the prior probabilities within those haplotypes as each animal's sequence is processed. Imputation of 1 million loci on 1 chromosome required 20 min and 5 gigabytes of memory using 10 processors for 500 bulls simulated at 8X coverage plus 250 younger bulls that had lower coverage or had high, medium, or low density chips. Percentages of correct genotypes were 99.2, 97.0, and 94.1 for bulls sequenced at 8X, 4X, and 2X coverage and were 98.1, 96.8, and 91.7 for bulls genotyped with 600K, 60K, and 10K density chips. Imputation using sequence with low coverage or high error was less accurate if genotypes from a high-density chip were not included in the sequence data.
Keywords: imputation; genotypes; sequence read depth; allele probabilities

Introduction

Sequence data with low read depth or high error requires different imputation strategies than previous algorithms designed for data from genotyping chips because homozygotes and heterozygotes are known with less certainty from sequence data (Menelaou and Marchini, 2012). Either of the 2 alleles may be read a different number of times, not read at all, or misread, which requires additional probability calculations not needed with present algorithms that assume genotypes are known. Animal breeders have developed imputation software that runs many times faster than software from human geneticists by using long-range phasing, general pedigrees, and the high degree of haplotypes shared across very large populations. For example, genotypes from several lower density chips are used to impute 60,000 markers for 500,000 animals in routine genomic evaluation and >600,000 markers for >150,000 animals in research (VanRaden et al., 2013).

These fast algorithms are adapted in the current report to use sequence data directly instead of first calling genotypes from the sequence reads. Within haplotype pairs, allele probabilities are updated based on read depth of the sequence alleles and prior probabilities. Including high-density (HD) chip genotypes for each sequenced animal ensures accurate haplotype matching for all animals. These strategies can reduce cost by accurately imputing genotypes using lower coverage for more individuals.

Materials and Methods

Simulated data. The population structure and number of sequence variants were identical to a previous test (VanRaden et al., 2013), but known genotypes were used in the previous test whereas read depth was reduced and errors were introduced in the current test. Sequence variants were simulated for a chromosome with a length of 1 Morgan and containing 1 million polymorphic loci which is equivalent to 30 million loci across the bovine genome. Sequence genotypes were then imputed from 20,000 markers/chromosome (simulating a 600K chip), 2,000 markers (simulating a 60K chip), or 333 markers (simulating a 10K chip). Allele frequencies in the founding population had a uniform distribution between 0 and 1 (u) for markers placed on chips, but the sequence variants had lower minor allele frequencies and a quadratic distribution, but with perhaps fewer rare variants than actual sequences.

Sequences were simulated for 250, 500, or 1,000 bulls that had the most US daughters, plus sequences reduced to 600K density for 250 randomly chosen US bulls born in 2010. Several generations sometimes separated the sequenced older bulls from the young bulls with 600K genotypes because the average birth year of the older bulls was 1987. Another 23,656 ancestors in the pedigree had sequences generated but not observed. Imputation accuracy was computed by percentage of matching true and called genotypes across all loci and also by the correlation after centering both the true and the called genotypes by subtracting twice the allele frequency.

True genotypes were reduced to read depths averaging 16, 8, 4, 2, or 1, whereas the previous test had assumed perfectly known genotypes for all sequence variants. The sequenced bulls had reduced read depth either at all variants or at all variants except the 600K markers on the chip, which were given a read depth of 32. This tested if genotypes should combine reduced-depth sequence data and an HD chip to improve imputation as recommended by Menelaou and Marchini (2012). Tests were repeated with 0, 1, 4, or 16% error in the individual sequence reads. Another test used reduced-depth sequence reads for the young bulls as proposed by De Donato et al. (2013) and by Hickey (2013), but these also included 600K chip data.

Human sequence genotypes for 394,724 SNPs from the shortest chromosome (HSA 22) from the 1000 Genomes Project Consortium (2012) were also imputed from 39,440 SNPs that had highest minor allele frequency (>0.12). The 1,092 individuals were randomly split into 884 reference sequences and 218 reduced to HD for validation.

Haplotype probabilities. Probability formulas and computer algorithms were derived and tested on the simulated bovine and actual human sequence data. Phasing of known genotypes into haplotypes may be simple if parents and progeny are genotyped but somewhat more difficult if the genotyped individuals are less related. Imputing missing genotypes can be done simply by

choosing the most frequent haplotype that does not conflict with the genotype, then obtaining the complement haplotype by subtracting known alleles in the first haplotype from the genotype, then searching for the next most frequent haplotype that agrees with the complement, and then filling any missing alleles in the 2 haplotypes if the genotype is homozygous (VanRaden et al., 2011).

Repeated application of inverse probability rules (Bayes theorem) can update haplotype probabilities with the new information provided by each individual's sequence data (S). The steps are analogous to those used in some long-range phasing algorithms. The most likely haplotype (H1) and its complement (H2) are selected from a list, or a new haplotype is added if none in the list are likely. Then posterior probabilities that the alleles h1 and h2 at a particular locus within H1 and H2, respectively, contain allele A are calculated from their prior probabilities (p_1 and p_2) and the individual's sequence data at that locus (s). This updating process is repeated for each individual using the posterior probabilities in H1 and H2 as prior probabilities for the next individual containing either of those same haplotypes. This accumulates linkage information into the haplotype list instead of using multi-locus math to account for linkage as in Duitama et al. (2011).

The sequence data s are coded simply as the numbers of A (n_A) and B (n_B) alleles observed; the 3 main categories of observed data were only n_A, only n_B, or both n_A and n_B positive. With no sequencing error, the third category always indicates a heterozygote, but the first 2 categories do not always indicate homozygotes because heterozygotes also produce only n_A or only n_B observations at a rate of $0.5^{(n_A + n_B)}$ each. For example, $n_A = 4$ and $n_B = 0$ could result from an AA homozygote producing A alleles every time or an AB heterozygote producing only A alleles with frequency of $0.5^4 = 0.0625$. With low-coverage sequence, n_A and n_B may both equal 0 at many loci, and those are treated as missing observations. Storage of n_A and n_B is more efficient than storing 3 genotype probabilities.

Prior probabilities that the 2 haplotypes contain an A at a particular locus are $P(\text{h1} = \text{A})$ and $P(\text{h2} = \text{A})$, and both initially are set to allele frequencies before processing the first individual. Posterior probabilities $P(\text{h1} = \text{A}|n_A,n_B)$ and $P(\text{h2} = \text{A}|n_A,n_B)$ are then obtained by jointly accounting for the 2 prior probabilities and using the standard inverse probability rule such as in Duitama et al. (2011):

$$P(\text{h1} = \text{A}|n_A,n_B) = P(n_A,n_B|\text{h1} = \text{A})[P(\text{h1} = \text{A})/P(n_A,n_B)];$$
$$P(\text{h2} = \text{A}|n_A,n_B) = P(n_A,n_B|\text{h2} = \text{A})[P(\text{h2} = \text{A})/P(n_A,n_B)].$$

Posterior probabilities accounted for the error in individual sequence reads (errate) using math similar to Druet et al. (2014), except that the probabilities were applied directly to haplotypes instead of first calling genotypes. For efficiency, probabilities of observing n_A and n_B given the 3 genotypes were calculated and stored AAprob, BBprob, and ABprob for later use when processing each potential haplotype. Factorial terms in the binomial distribution were not computed because they always canceled in the likelihood ratios:

$$\text{AAprob} = P(n_A,n_B|\text{AA}) = \text{errate}^{n_B}\,(1 - \text{errate})^{n_A};$$
$$\text{BBprob} = P(n_A,n_B|\text{BB}) = \text{errate}^{n_A}\,(1 - \text{errate})^{n_B};$$
$$\text{ABprob} = P(n_A,n_B|\text{AB}) = 0.5^{(n_A + n_B)}.$$

The 2 prior probabilities $P(\text{h1} = \text{A})$ and $P(\text{h2} = \text{A})$ were labeled p_1 and p_2 for simplicity. From these, the conditional probability of observing n_A and n_B given that h1 contains A or that h2 contains A were calculated from

$$P(n_A,n_B|\text{h1} = \text{A}) = p_2\,\text{AAprob} + (1 - p_2)\,\text{ABprob};$$
$$P(n_A,n_B|\text{h2} = \text{A}) = p_1\,\text{AAprob} + (1 - p_1)\,\text{ABprob}.$$

The unconditional probability of observing n_A and n_B was computed by summing probabilities that the true genotype was AA, BB, or AB multiplied by the probability of observing n_A and n_B given each genotype. The overall probability for the population used the same formula except that population frequency p was substituted for the haplotype prior probabilities p_1 and p_2:

$$P(n_A,n_B) = p_1p_2\,\text{AAprob} + (1 - p_1)(1 - p_2)\,\text{BBprob} + (p_1 + p_2 - 2p_1p_2)\,\text{ABprob}.$$

The H1 and H2 mostly likely to form the genotype were selected using likelihood ratio tests from a haplotype list that was sorted by descending frequency. The probability of observing s at each locus was divided by the probability that s would be observed if alleles were chosen randomly from the population ($p_2 = p$), and these ratios at each locus in a potential H1 were multiplied to obtain the joint likelihood ratio $P(\text{S}|\text{H1})/P(\text{S})$. A particular haplotype H1 was selected if the joint likelihood ratio was $>1/n$, where n is the number of loci with observed data in the haplotype. The H2 was selected if the joint likelihood of S given H2 and H1 divided by the likelihood given H1 [i.e., $P(\text{S}|\text{H1},\text{H2})/P(\text{S}|\text{H1})$] was $>1/\{n[1 + (n/100)]\}$. Also, if the likelihood ratio was $<1/n$ at any individual locus, the haplotype was discarded immediately to save computation.

The 2 selected H1 and H2 were updated by combining the formulas above to obtain their posterior probabilities of containing allele A given the sequence data and the 2 haplotype prior probabilities at each locus:

$$P(\text{h1} = \text{A}|n_A,n_B) = [p_2\,\text{AAprob} + (1 - p_2)\,\text{ABprob}]\,p_1/[p_1p_2\text{AAprob} + (1 - p_1)(1 - p_2)\,\text{BBprob} + (p_1 + p_2 - 2p_1p_2)\,\text{ABprob}];$$
$$P(\text{h2} = \text{A}|n_A,n_B) = [p_1\,\text{AAprob} + (1 - p_1)\,\text{ABprob}]\,p_2/[p_1p_2\text{AAprob} + (1 - p_1)(1 - p_2)\,\text{BBprob} + (p_1 + p_2 - 2p_1p_2)\,\text{ABprob})].$$

Results and Discussion

Imputation of 1 million SNP on 1 chromosome for 1,000 animals took 20 min using 10 processors and 5 gigabytes of memory. Time required is only slightly more than findhap (VanRaden et al., 2011), but memory is almost twice because of storing 2-byte probabilities instead of 1-byte haplotype codes. When imputing from HD to sequence using 16X coverage for 500 bulls, 98.4% of genotypes were correct with the new algorithm compared with 97.8% with version 2 of findhap (VanRaden et al., 2013). Slight improvements compared to the squared correlations from

Beagle in Druet et al. (2014) also indicate that updating haplotype probabilities may be a more accurate strategy with high read depth and also allows including sequences with low read depth (Table 1). With 4X instead of 16X coverage, 97.0% of the imputed genotypes were correct for the sequenced bulls and 96.7% for bulls with 600K. High read depth is desired if the goal is direct investigation of the sequenced bulls, but lower depth is desired if the goal is imputation from HD. Sequencing twice as many bulls at half the read depth often gave more correct calls (e.g., 95.3% using 250 bulls with 16X, 98.1% using 500 bulls with 8X, or 98.2% using 1,000 bulls with 4X).

Table 1. Correctly imputed genotypes (%) and correlations (R) by read depth and number of bulls with sequence data containing 1% error and high-density (HD) markers

	Read	From sequence		From HD markers	
Bulls	depth	R	Correct	R	Correct
1,000	16	0.999	99.9	0.989	99.2
	8	0.992	99.5	0.987	99.1
	4	0.969	97.8	0.975	98.2
	2	0.942	95.9	0.951	96.5
	1	0.912	93.8	0.911	93.8
500	16	0.999	99.9	0.977	98.4
	8	0.988	99.2	0.974	98.1
	4	0.958	97.0	0.954	96.7
	2	0.919	94.1	0.917	94.1
250	16	0.998	99.9	0.931	95.3
	8	0.981	98.7	0.926	95.0
	4	0.939	95.8	0.897	93.1

Sequence reads with 1 or 4% error rates reduced imputation success slightly, but 16% error caused larger reduction (Table 2). For example, with 8X coverage, the 500 sequenced bulls had 99.5% correct calls if sequence error was 0%, 99.2% if 1%, 98.4% if 4%, and 95.2% if 16% error. The HD bulls had 98.2, 98.1, 98.0, and 96.9% correct calls, respectively. With lower read depths, error rates caused larger differences in imputation success, but with 1,000 bulls and 16X, >99% of genotypes for both sequenced and HD bulls were called correctly even if the read error rate was 16%. Exclusion of HD chip data for the sequenced bulls reduced accuracy with low read depth or high error, which was consistent with the results of Menelaou and Marchini (2012). For example, with 8X coverage and 4% error, imputation success dropped from 98.4 to 97.8% for sequenced bulls and from 98.0 to 97.2% for HD bulls (Table 2) because detecting common haplotypes was difficult if the sequenced bulls had few data at the HD loci. Results were similar for HSA 22 but with lower maximums of 97.4% correct and correlation of only 86% when imputing from HD because of rare SNPs and lower linkage disequilibrium in the human sequence data.

Bull genotypes from less dense chips had 96.8% correct imputation from 60K or 91.7% from 10K chips when 1,000 bulls had sequence data with 1% error, read depth 8, and HD markers included. To include sufficient markers within each interval, maximum haplotype length was extended to 100,000 from the 50,000 used for imputing from 600K. When sequences with lower read depth were included, bulls had 99.1% correct from 2X coverage or 99.0% correct from 1X coverage. A final analysis included bulls with 8X, 2X, 1X, 600K, 60K, and 10K all in 1 data set, but imputation accuracy was slightly reduced in some cases because haplotype lengths and other options were not optimal for all bulls. Further research may improve these strategies and the resulting accuracy of fast imputation.

Table 2. Correctly imputed genotypes (%) and correlations (R) by error rate using 500 bulls with sequence data with read depth 8 that included or excluded high-density (HD) markers

HD	Error	From sequence		From HD markers	
included?	rate	R	Correct	R	Correct
Yes	0.00	0.994	99.5	0.975	98.2
	0.01	0.988	99.2	0.974	98.1
	0.04	0.977	98.4	0.972	98.0
	0.16	0.932	95.2	0.956	96.9
No	0.00	0.991	99.4	0.973	98.1
	0.01	0.985	98.9	0.970	97.9
	0.04	0.969	97.8	0.961	97.2
	0.16	0.911	93.8	0.874	91.5

Conclusion

Genotypes can be imputed more accurately by using each animal's raw sequence reads to update allele probabilities within pairs of haplotypes when simulated sequences have high error rates or medium to low coverage. Sequencing tools offer a tradeoff between number of animals and average read depth. More efficient imputation will allow geneticists to locate and test effects of more DNA variants and to include those in future selection programs.

Literature Cited

De Donato, M., Peters, S. O., Mitchell, S. E. et al. (2013). PLoS ONE 8:e62137.

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Heredity 112:39–47.

Duitama, J., Kennedy, J., Dinakar, S. et al. (2011). BMC Bioinformatics 12(Suppl. 1):S53.

Hickey, J. M. (2013). J. Anim. Breeding Genet. 130:331–332.

Menelaou, A., and Marchini, J. (2013). Bioinformatics 29:84–91.

The 1000 Genomes Project Consortium. 2012. Nature 491: 56–65.

VanRaden, P. M., Null, D. J., Sargolzaei, M. et al. (2013). J. Dairy Sci. 96:668–678.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R. et al. (2011). Genet. Sel. Evol. 43:10.