SEQUENCING STRATEGY, IMPUTATION AND GENOMIC PREDICTION IN A LARGE PIG SEQUENCING STUDY

R. Ros-Freixedes[1,2], A. Whalen[1,] A. Somavilla[1], S. Gonen[1], M. Battagin[1], M. Johnsson[1,3], G. Gorjanc[1], C.Y. Chen[4], W.O. Herring[4], A.J. Mileham[5] and J.M. Hickey[1]

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush Campus, EH25 9RG Midlothian, Scotland, UK
[2]Departament de Ciència Animal, Universitat de Lleida-Agrotecnio Center, Lleida, Spain.
[3]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-750 07, Uppsala, Sweden
[4]Genus plc, 100 Bluegrass Commons Blvd., Suite 2200, Hendersonville, TN, 37075 USA
[5]Genus plc, 1525 River Road, DeForest, WI, 53532 USA

SUMMARY

The use of whole-genome sequence data has great potential in livestock breeding programs but suitable sequencing strategies and imputation methods need to be developed to generate sequence information for a large number of individuals at an affordable cost. We describe the sequencing strategy that we followed in a study that sequenced more than 7,848 pigs from nine commercial lines, mostly at low coverage. Results demonstrate that the coupling of appropriate sequencing strategies and imputation methods such as hybrid peeling is a viable strategy for producing whole-genome sequence data for large livestock pedigreed populations, but it remains to be determined whether these large datasets can provide an increased accuracy of genomic predictions.

INTRODUCTION

The use of whole-genome sequence data has great potential in livestock breeding programs. It may increase the power of discovery of causative variants (Pasanuic *et al.* 2012; Daetwyler *et al*. 2014; Nicod *et al.* 2016) and may enable more accurate and persistent predictions of breeding values than marker arrays (Meuwissen and Goddard, 2010; Iheshiulor *et al.* 2016). To capture the full potential of sequence data in livestock, sequence and phenotype data are required on a large number of individuals, perhaps millions, to accurately estimate the effects of the large number of causative variants that underlie quantitative traits (Hickey *et al*, 2014).

Low-cost sequencing strategies combined with imputation can be utilised to generate the required amount of sequence information for a large number of individuals at an affordable cost (Brøndum *et al.* 2014; van Binsbergen *et al.* 2014; VanRaden *et al.* 2015; Pausch *et al.* 2017). Low-coverage sequencing (LCSeq) enables the sequencing of a larger number of animals, which provides four advantages: (1) higher variant discovery rates, particularly for low-frequency variants; (2) inclusion of rare haplotypes; (3) a more precise capture of the recombination events that have occurred in the population, which enables better definition of haplotypes and thus better imputation of these haplotypes into the individuals that carry them; and (4) more sequenced animals that are related, which improves the imputation of the sequence data to the whole population.

We first describe the sequencing strategy that we followed in a study that sequenced more than 7,848 pigs from nine commercial lines, mostly at low coverage (1x or 2x). Then, we demonstrate that the coupling of that sequencing strategies with the imputation method 'hybrid peeling' is a viable strategy for producing whole-genome sequence data for large livestock pedigreed populations. Finally, we test the benefit that these large datasets can provide an increased accuracy of genomic predictions.

MATERIALS AND METHODS

Sequencing strategy. We performed whole-genome sequencing of 7,848 individuals from nine commercial pig breeding lines (Genus PIC, Hendersonville, TN) with a total coverage of approximately 32,114x. Sequencing effort in each of the nine lines was proportional to population size. Approximately 2% (1.7-2.5%) of the pigs in each line were sequenced. Most pigs were sequenced at low coverage, with target coverage of 1 or 2x, but a subset of pigs were sequenced at higher coverage of 5x, 15x, or 30x. Thus, the average individual coverage was 4.1x, but the median coverage was 1.5x.

We selected the individuals and the coverage at which they were sequenced using a three-step strategy: (1) we first selected sires and dams that contributed most genotyped progeny in the pedigree (referred to as 'top sires and dams') to be respectively sequenced at 2x and 1x; (2) conditional on the first step, we used AlphaSeqOpt part 1 (Gonen *et al.* 2017) to identify the individuals whose haplotypes represented the greatest proportion of the population haplotypes (referred to as 'focal individuals') and to determine an optimal level of sequencing coverage between 0x and 30x for these individuals and their immediate ancestors (i.e., parents and grandparents) under a total cost constraint; and (3) conditional on the second step, we used the AlphaSeqOpt part 2 (Ros-Freixedes *et al.*, 2017) to identify individuals that carried haplotypes whose cumulative coverage was low (i.e., below 10x) and distributed 1x sequencing amongst those individuals so that the cumulative coverage on the haplotypes could be increased (i.e., at or above 10x). AlphaSeqOpt used haplotypes inferred from marker array genotypes (GGP-Porcine HD BeadChip; GeneSeek, Lincoln, NE), which were phased with AlphaPhase (Hickey *et al.* 2011) and imputed with AlphaImpute (Hickey *et al.*, 2012). The sequencing resources were split so that approximately 30% of the sequencing resources were used for sequencing the top sires at 2x, 15% for the top dams at 1x, 25% for the focal individuals and their immediate ancestors at variable coverage, and the remaining 30% for individuals that carried under-sequenced haplotypes at 1x.

Variant discovery. The reads were preprocessed using Trimmomatic (Bolger *et al.* 2014) to cut adapter sequences from the reads. Then the reads were aligned to the Sscrofa11.1 reference genome using the BWA-MEM algorithm (Li & Durbin 2009). Duplicates were marked with Picard (http://broadinstitute.github.io/picard). SNPs and short insertions and deletions (indels) were genotyped jointly for all samples using a pipeline based on the HaplotypeCaller tool from GATK 3.8 (DePristo *et al.* 2011). To avoid biases towards the reference allele introduced by GATK when applied on low-coverage sequence data we extracted the read counts supporting each allele directly from the aligned reads stored in the BAM files with a pile-up function using the pipeline described in (Ros-Freixedes *et al.* 2018). A total of 60 million SNPs were discovered across the nine lines.

Imputation of whole-genome sequence data. Most individuals in every population were genotyped using commercial marker arrays, with either 15,000 (LD) or 75,000 (HD) markers genome-wide. Imputation to whole-genome sequence was performed in each population separately using hybrid peeling, as implemented in AlphaPeel (Whalen *et al*. 2018) with the default settings. This method involves two stages: (1) multi-locus iterative peeling to estimate the segregation (the probability that each pair of grandparental gametes was co-inherited at a given locus) at the positions genotyped with the marker arrays; and (2) a modified single-locus iterative peeling step to impute the genotypes at each variant position discovered from the sequence data. This two-stage method reduces the computational cost of the imputation by estimating segregation of the markers from the array only and then approximating the segregation estimates at any other loci based on the estimates of the markers from the array that flank them. The accuracy loss of this approximation is negligible due to the limited number of recombinations in each chromosome and the high probability that nearby markers are inherited together. Multi-locus iterative peeling was performed on all available marker array data to estimate the segregation probabilities for each individual. The individuals genotyped

with LD marker arrays were not imputed to HD prior to this step. The segregation probabilities were used for segregation-aware single-locus iterative peeling for the remaining segregating variants. The total number of pigs with imputed data across the nine lines ascends to around 350,000.

To assess imputation accuracy, we used 284 individuals from four of the nine populations who were sequenced at high coverage (15x or 30x). Of these, 37 belonged to a 20,000-individual population, 65 to a 35,000-individual population, 92 to a 70,000-individual population, and 90 to a 110,000-individual population. Many of these individuals sequenced at high coverage belonged to early generations of the pedigree of each population. Sequence data of the 284 individuals was completely masked, using a leave-one-out design. The imputed allele dosages were compared to those obtained with the complete data, considered as the 'true' values. For estimating the accuracies, we used 50,000 non-consecutive SNPs chosen randomly from chromosome 5.

Genomic prediction. Genomic prediction accuracy was tested in a single line with 30k pigs with imputed genotypes for 16 million of SNPs. Genomic predictions were performed using ridge regression as implemented in AlphaBayes software. The model was trained on 22,318 individuals and validated on 1,458 individuals. Genomic predictions were performed for nine synthetic traits with different heritability (0.1, 0.25, or 0.5) and with different number of QTN underlying their variation (100, 1,000, or 10,000 QTN), selected randomly from among all variants. The effect of the QTN was sampled from a normal distribution $N(0,1)$. Genomic predictions were performed using four sets of markers: the 57k markers from the array (HD), 248k variants preselected from the sequence data based on LD pruning (WGS_LD), around 183k variants preselected from the sequence data based on results of single-marker regression with a set of 13k individuals independent from the training and testing sets (WGS_SMR), or 67k variants preselected from the sequence data by keeping only every 200th variant (WGS_200th). Accuracy of the gEBV was estimated as the correlation between the gEBV and the synthetic phenotypes in the validation set.

RESULTS AND DISCUSSION

Imputation accuracy. The imputation accuracy in the real data was high for most of the tested individuals. The imputation accuracy achieved for each of the 284 tested individuals is shown in Figure 1. The average individual-wise dosage correlation was 0.94 but there was substantial variation with an asymmetrical distribution (median: 0.97; min: 0.11; max: 1; interquartile range: 0.94-0.98). Some of the oldest individuals that belonged to the earliest generations of the pedigree (some of the 106 individuals located in the first 20% of the pedigree) had lower imputation accuracy than individuals in the remainder of pedigree, who had consistently high imputation accuracy. This pattern was observed for all four populations. The imputation accuracy of the individuals in later generations (the 178 individuals after the first 20% of the pedigree) was higher, with an average dosage correlation of 0.97 and with much lower variability (median: 0.98; min: 0.69; max: 1; interquartile range: 0.96-0.99).

The marker array density of the individuals was confounded with the number of ancestors that were genotyped with marker arrays. The non-genotyped individuals (n=19) and approximately half of the individuals genotyped at HD (n=87 out of 157) belonged to early generations of the pedigree, which reduced the chances that they had ancestors with data and penalized the imputation accuracy for these two groups of individuals. On the contrary, most individuals genotyped at LD belonged to later generations (n=91 out of 108), ensuring that their ancestors had enough data to enable high imputation accuracies for the LD individuals. The average dosage correlation for the non-genotyped individuals was 0.81, for the HD individuals was 0.94, and for the LD individuals was 0.96. The average dosage correlation for the HD individuals in the earliest generations was lower (0.91) than for the HD individuals in later generations (0.97). For individuals in the later generations there were no significant differences between marker array densities and the average dosage correlation of both

the HD and LD individuals was 0.97 and therefore no intermediate imputation steps were required for the LD individuals. There was no clear trend that population size affected imputation accuracy.

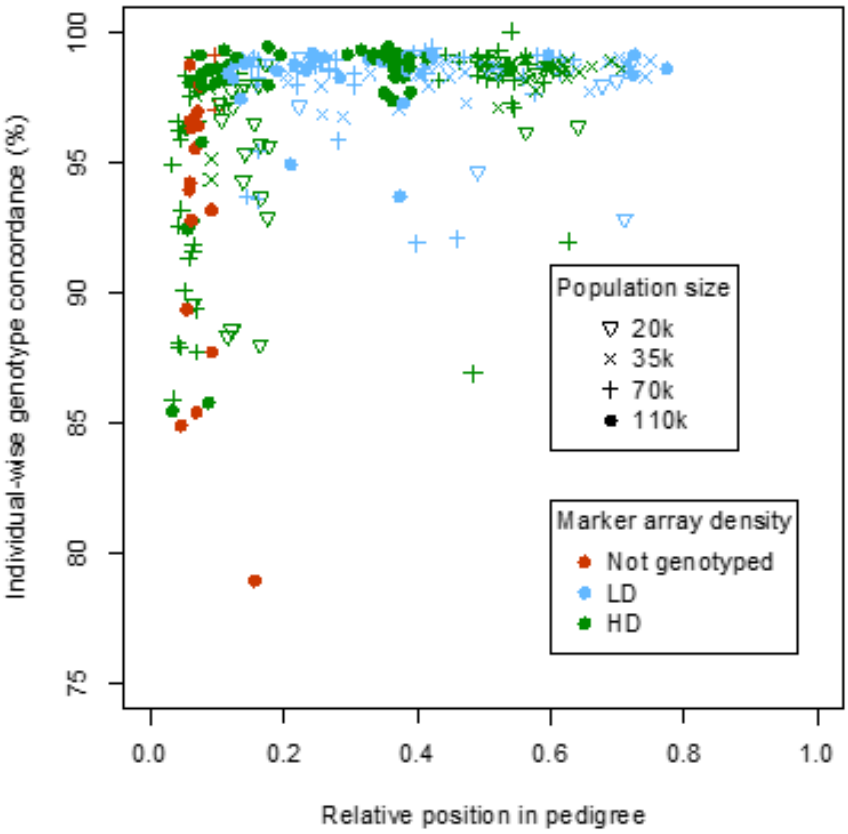

Figure 1. Imputation accuracy on relative position of the individual in the pedigree, marker array density, or population size

Genomic prediction. Sequence data can provide better prediction accuracy than marker arrays in some cases, but its advantage may depend on the genetic architecture of the trait. The genomic prediction accuracies for the nine synthetic traits are shown in Table 1. When a low number of QTN determine the phenotype, there may be sufficient statistical power to identify variants that underlie the genetic variation of the trait and prediction accuracy using those variants (WGS_SMR) is higher than with the markers from commercial marker arrays (HD). This is consistent with previous observations that adding one or a few markers with large effects as predictors can improve prediction accuracy of the marker arrays (Estany *et al*. 2017; Lopes *et al*. 2017; Nani *et al*. 2019; Al Kalaldeh *et al*. 2019). In such contexts, the information from markers with large effect could overcome the noise that arises from a higher number of markers with low effects. When the number of QTN is large, it became more difficult to identify these variants with single-marker regression and WGS_SMR performed worse than HD. In such cases, other sets of variants selected from the sequence data can be (marginally) more beneficial than the commercial marker arrays as they are not affected by ascertainment bias in the same way as commercial marker arrays.

Table 1. Prediction accuracies for nine synthetic traits

QTN	h^2	HD	WGS_LD	WGS_SMR	WGS_200th
100	0.1	0.370	0.367	0.389	0.368
	0.25	0.416	0.395	0.422	0.418
	0.5	0.625	0.615	0.626	0.626
1,000	0.1	0.373	0.345	0.356	0.370
	0.25	0.396	0.393	0.402	0.404
	0.5	0.620	0.594	0.597	0.620
10,000	0.1	0.430	0.411	0.395	0.430
	0.25	0.437	0.430	0.398	0.444
	0.5	0.657	0.644	0.617	0.658

In this test we did not observe an improvement in prediction accuracy using sequence data when the number of QTN was large, which is the case of many traits of economic interest in livestock. These results are partly due to the already high prediction accuracies obtained with the current implementation of genomic selection using commercial marker arrays. These results are in line with other studies that found no improvement or only small variations in genomic prediction when using sequence data, often by preselecting variants, compared to HD marker arrays (van Binsbergen *et al*. 2015; Calus *et al*. 2016; Veerkamp *et al*. 2016; van den Berg *et al*. 2017; VanRaden *et al*. 2017). However, these genomic prediction results are preliminary results for a single line. With a more complete set of sequenced individuals, it remains to be determined whether the results will improve due to: data from multiple breeds, enabling multi-breed training and a much larger training set; or genomic prediction methods that are more suited for exploiting sequence data at a large scale than ridge regression.

CONCLUSIONS

The coupling of an appropriate sequencing strategy and hybrid peeling is a powerful method for generating whole-genome sequence data in large pedigreed populations, as long as the individuals are connected to enough informative relatives with marker array or sequence data, and regardless of population size. It remains to be determined whether these large datasets can provide the leverage for increased accuracy of genomic predictions.

REFERENCES

Al Kalaldeh M., Gibson J., Duijvesteijn N., Daetwyler H.D., MacLeod I., Moghaddar N., Lee S.H. and van der Werf J.H.J. (2019) *Genet. Sel. Evol.* **51**: 32.

Bolger A.M., Lohse M. and Usadel B. (2014) **30**: 2114.

Brøndum R.F., Guldbrandsten B., Sahana G., Lund M.S. and Su G. (2014) *BMC Genomics* **15**: 728.

Calus M.P.L., Bouwman A.C., Schrooten C. and Veerkamp R.F. (2016). *Genet. Sel. Evol*. **48**: 49.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., van Binsbergen R., R.F. Brøndum R.F., …and Hayes B.J. (2014) *Nat. Genet*. **46**: 858.

DePristo M.A, Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., …and Daly M.J. (2011) *Nat. Genet*. **43**: 491.

Estany J., Ros-Freixedes R., Tor M. and Pena R.N. (2017) *J. Anim. Sci.* **95**: 2261.

Gonen S., Ros-Freixedes R., Battagin M., Gorjanc G. and Hickey J.M. (2017) *Genet. Sel. Evol*. **49**: 47.

Hickey J.M., Kinghorn B.P., Tier B., Wilson J.F., Dunstan N. and van der Werf J.H.J. (2011) *Genet Sel Evol*. **43**: 12.
Hickey J.M., Kinghorn B.P., Tier B., van der Werf J.H.J. and Cleveland M.A. (2012) *Genet Sel Evol*. **44**: 9.
Hickey J.M., Gorjanc G., Cleveland M.A., Kranis A., Jenko J., Mészáros G., Woolliams J.A. and Perez-Enciso M. (2014) *Proc. 10th World Congress of Genetics Applied to Livestock Production*, 17-22 Aug, Vancouver, BC, Canada.
Iheshiulor O.O.M., Woolliams J.A., Yu X., Wellmann R. and Meuwissen T.H.E. (2016) *Genet. Sel. Evol*. **48**: 15.
Li H. and Durbin R. (2009) *Bioinformatics* **25**: 1754.
Lopes M.S., Bovenhuis H., van Son M., Nordbø Ø., Grindflek E.H., Knol E.F. and Bastiaansen J.W. (2017) *J. Anim. Sci.* **95**: 59.
Meuwissen T. and Goddard M. (2010) *Genetics* **185**: 623.
Nani J.P., Rezende F.M. and Peñagaricano F. (2019) *BMC Genomics* **20**: 258.
Nicod J., Davies R.W., Cai N., Hassett C., Goodstadt L., Cosgrove C., …and Flint J. (2016) *Nat. Genet*. **48**: 912.
Pasanuic B., Rohland N., McLaren P.J., Garimella K., Zaitlen N., Li H., …and Price A.L. (2012) *Nat. Genet*. **44**: 631.
Pausch H., MacLeod I.M., Fries R., Emmerling R., Bowman P.J., Daetwyler H.D. and Goddard M.E. (2017) *Genet. Sel. Evol*. **49**: 24.
Ros-Freixedes R., Battagin M., Johnson M., Gorjanc G., Mileham A.J., Rounsley S.D. and Hickey J.M. (2018) *Genet. Sel. Evol.* **50**: 64.
Ros-Freixedes R., Gonen S., Gorjanc G. and Hickey J.M. (2017) *Genet. Sel. Evol.* **49**: 78.
van Binsbergen R., Bink M.C.A.M., Calus M.P.L., van Eeuwijk F.A., Hayes B.J., Hulsegge I. and Veerkamp R.F. (2014) *Genet. Sel. Evol*. **46**: 41.
van Binsbergen R., Calus M.P.L., Bink M.C.A.M., van Eeuwijk F.A., Schrooten C. and Veerkamp R.F. (2015) *Genet. Sel. Evol*. **47**: 71.
van den Berg I., Bowman P.J., MacLeod I.M., Hayes B.J., Wang T., Bolormaa S. and Goddard M.E. (2017) *Genet. Sel. Evol*. **49**: 70.
VanRaden P.M., Sun C. and O'Connell J.R. (2015) *BMC Genet*. **16**: 82.
VanRaden P.M., Tooker M.E., O'Connell J.R., Cole J.B. and Bickhart D.M. (2017) *Genet. Sel. Evol.* **49**: 32.
Veerkamp R.F., Bouwman A.C., Schrooten C. and Calus M.P.L. (2016) *Genet. Sel. Evol.* **48**: 95.
Whalen A., Ros-Freixedes R., Wilson D.L., Gorjanc G. and Hickey J.M. (2018) *Genet Sel Evol.* **50**: 67.