Imputation accuracy of whole-genome sequence in pigs

A.C. Bouwman¹, M. van Son², B. Harlizius³ & B. Zambach³

¹ Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands
Aniek.Bouwman@wur.nl (Corresponding Author)
² Topigs Norsvin, Storhamargata 44, 2317 Hamar, Norway
³ Topigs Norsvin Research Centre, PO Box 43, 6640 AA Beuningen, the Netherlands

Summary

In the current study we evaluated imputation accuracy for Sus scrofa autosome 7 (SSC7) for imputation from a 70k SNP chip to whole-genome sequence (WGS), from a 660k SNP chip to WGS, and a two-step procedure from 70k to 660k to WGS, using Beagle 4.1, Minimac3 and FImpute. A reference population of 168 pigs with WGS data from several European breeds and lines was used to do a leave-one-out cross-validation. Results showed that imputation from 660k is more accurate compared to imputation from 70k directly to WGS. The two-step procedure from 70k to 660k to WGS showed the lowest imputation accuracy due to relatively poor imputation accuracy in the first step due to the limited reference population. Minimac3 seemed to perform best, but with a closer look at the details, Beagle 4.1 outperformed Minimac3. FImpute performed less well, but was not used to its full potential here as pedigree information was not included in the imputation. Imputation from commercial SNP chips to WGS in pigs is possible, in practise with even better results than shown here, however imputation accuracy is limited by the size of the current reference population at hand.

Keywords: accuracy, imputation, pig, whole-genome sequence

Introduction

To be able to use whole-genome sequence (WGS) in quantitative genetics a lot of individuals with WGS are required. Currently, it is economically unfeasible to sequence whole populations, hence imputation from chip genotypes to WGS using a sequenced reference population is a cost-effective alternative. Imputation to WGS in order to do GWAS or genomic prediction has been applied for instance in cattle (Binsbergen et al., 2015, Sanchez et al., 2017, Wang et al., 2017). In cattle, the 1000 Bull Genomes consortium (Daetwyler et al., 2014) has enhanced the use of WGS by sharing sequenced individuals among partners. In pig breeding the number of sequenced animals is lagging behind because it is mainly within-company data. However, imputation to WGS is of interest to the entire pig breeding industry and for research.

A lot of QTL and significant SNP from GWAS have been detected for economically relevant traits in pork production. One of the regions of interest is located on Sus scrofa autosome 7 (SSC7) and is associated with the number of teats (Lopes et al., 2014), as well as vertebral number in pigs (Fan et al., 2013). These are correlated traits: more vertebra leads to more space for teats, and with more teats a sow can nurture more piglets. It is therefore an interesting region to investigate for causal mutation, as there is potential to boost selection for number of teats and enhance traits such as number of piglets weaned.
In order to fine-map this QTL region, imputation to WGS would be needed. The question is how accurate such imputation is with current numbers of sequenced reference individuals. Although evaluation of imputation accuracy from commercial SNP chips to WGS has been assessed in several species, the number of studies in pigs is minimal. To our knowledge only Yan et al. (2017) briefly addressed the topic and evaluated imputation from a 60k SNP chip to WGS using Beagle 4.1.

In our current study we evaluated imputation accuracy for SSC7 for imputation from a 70k SNP chip to WGS, from a 660k SNP chip to WGS, and a two-step procedure from 70k to 660k to WGS, using Beagle 4.1, Minimac3 and FImpute.

Material and methods

Population

A reference population of 168 pigs with WGS data was used to evaluate imputation accuracy. The reference population contains pigs from several breeds and lines from both Norway and the Netherlands (Topigs Norsvin). The population consisted of 70 Large White, 46 Duroc, 37 Landrace, 13 Pietrain, 1 Large White-Landrace crossbred, and 1 animal with unknown breed. Four animals had one parent sequenced in the reference population. Variant calling and imputation were carried out for Sus scrofa autosome 7 (SSC7) because this chromosome harbours an interesting region associated with the number of teats (Lopes et al., 2014), as well as vertebral number in pigs (Fan et al., 2013), which are correlated traits.

Variant calling

Raw sequence data was aligned to the pig genome build Sscrofa10.2 (Ensembl72) (Groenen et al., 2012) using Samtools (Li et al., 2009). Variants (SNP and InDels) were called with GATK unified genotyper (McKenna et al., 2010) using all 168 samples. Default settings were used, but in addition stand_call_conf was set to 30.0, stand_emit_conf was set to 20.0 and dco = 200. Variants were filtered using VCFtools (Danecek et al., 2011) with the following criteria: read depth values (per individual) ≥ 4 and ≤ 35, overall PRED Quality score above 20, variants with more than 20% missing data were excluded, only bi-allelic variants were considered, and sites were thinned so that no two sites were within 3bp from one another. After variant filtering, the reference population was phased and missing genotypes in the sequence data were imputed using Beagle 4.1 with 10 phasing iterations (Browning & Browning, 2007).

Imputation

Imputation of SSC7 from two commercial SNP-chips, Porcine GGP-HD with ~70k SNP (GeneSeek, Lincoln, NE), and PigHD Axiom with ~660k SNP (Affymetrix, Santa Clara, CA), to WGS was evaluated. Imputation from 70k to WGS, as well as, imputation in two steps from 70k to 660k to WGS was evaluated with Beagle 4.1. For imputation from 660k to WGS we evaluated three imputation programs. The following imputation programs were used with the described settings: i) Beagle 4.1 (Browning & Browning, 2016) default settings, but with 10 iterations and Ne set to 1,000; ii) FImpute (Sargolzaei et al., 2014) default settings, but without pedigree; iii) Minimac3 (Das et al., 2016) default settings.

A leave-one-out cross-validation was performed imputing one animal whose sequence
variants were masked leaving only the chip variants as lower density genotypes and the other 167 animals as a sequenced reference population. The imputation accuracy for each variant was assessed as the Pearson correlation coefficient \( r \) between true and imputed genotypes. The true genotypes were known since all individuals were sequenced and the sequence variants not on the chip were masked. Both Beagle 4.1 and Minimac3 provided allele dosages (DS) as well as most likely genotypes (ML), hence for those two imputation programs the imputation accuracy of both genotype coding methods were assessed. Only variants with at least 6 alleles (MAF>0.035 in this case) segregating in the population were considered in the evaluation of imputation accuracy.

**Results**

The program GATK detected 2,529,312 sites on SSC7 of which 1,664,543 were kept after filtering. In total, 3,356 variants from the 70k chip, and 29,249 variants from the 660k chip, were detected in the WGS of SSC7 (and thus segregating in the 168 animals).

Mean imputation accuracies in terms of correlation are provided in Table 1. Imputation from 660k to WGS was more accurate than imputation from 70k. Imputing in two steps from 70k to 660k to WGS did not improve imputation accuracy in this case. The imputation accuracy in the first step from 70k to 660k was 0.839. For imputation of 660k to WGS with different imputation programs the accuracy of imputation ranged between 0.742 and 0.889. Minimac3 ML showed the highest imputation accuracy, however, based on a lower number of variants (Table 1). Comparing only overlapping variants (i.e. imputed variants for which it was possible to calculate a correlation with the true genotype due to variation (polymorphic) in the imputed genotypes), shows that Beagle 4.1 performs best and that dosages are preferred above most likely genotypes.

Figure 1 shows the relationship between imputation accuracy and MAF. The figure indicates that imputation of low MAF variants remains challenging. Plotting the imputation accuracy against the position on SSC7 shows that there are some poorly imputed regions (Figure 2). For some regions, this is most likely because the density of the 660k chip was low, e.g., the region around 26Mbp.

**Discussion**

A two-step procedure did not improve imputation accuracy to WGS, which was inconsistent with findings from another cattle study (Binsbergen et al., 2014). In practise, it might be beneficial to impute in two steps, because there will be far more animals with 660k genotypes, and they will be more related to the individuals to impute than the 167 sequenced animals used as a reference population for imputation to 660k here. This would result in higher imputation accuracy in the first step from 70k to 660k, and therefore also in higher overall imputation accuracy compared to direct imputation from 70k to WGS. From Table 1 it can be assumed that if imputation accuracy in the first step is higher than 0.874 (0.708/0.810), a two-step imputation procedure improves the imputation accuracy from 70k to WGS (using Beagle 4.1 ML).

Minimac3 ML had the highest imputation accuracy. However, to some extent this comparison between imputation programs is not fair. The imputation accuracy was assessed as the correlation between true and imputed genotypes, but for variants imputed as monomorphic there is no variation to actually calculate the correlation. Hence variants falsely imputed as monomorphic were not accounted for in the mean accuracy. From Table 1 it can
be seen that the number of variants included for the calculation of the mean imputation accuracy, for Minimac3 ML is much lower than for the other imputation methods. This indicates that the remaining variants were imputed as monomorphic while there were at least 6 alleles segregating in the population under investigation, so at least 4 alleles segregating in the reference population. This is mainly the case for low MAF variants, which generally have a lower imputation accuracy, and therefore the results for Minimac3 ML are overestimated. Also, because this mainly occurs in low MAF variants, it is an issue when assessing imputation accuracy for WGS data. With imputation to commercial SNP chips, where SNP were selected based on their common frequency in the commonly used breeds, this is not much of an issue. Comparing imputation accuracy over all (859,296) variants that received an imputation accuracy with all imputation methods, Beagle 4.1 DS performs best (Table 1).

The imputation program FImpute was not used to its full potential as FImpute has the advantage to be able to use both full pedigree information and population information. Here, the sequenced reference population consisted of founder animals of several lines and breeds, with limited relationship among each other. In practise, the set of sequenced animals are often the founders of the population to impute. Therefore FImpute might actually perform better than it did in the current study.

Imputation success is dependent on the quality of the genome build. For the pig genome build Sscrofa 10.2 it is known that there are many errors in the reference genome. Recently, pig assembly Sscrofa11.1 was released, which has improved considerably with respect to the previous build (personal communication M.A.M. Groenen). Therefore imputation accuracy for SSC7 using at least one of the imputation programs will be assessed again after aligning the sequence data of this reference population to Sscrofa 11.1.

Compared to our study, Yan et al. (2017) detected more variants on SSC7 in their sequenced population because they had a larger set of animals from a wider range of pig breeds including Asian breeds and wild boar. The average imputation accuracy of 0.80 for imputation from Illumina PorcineSNP60 using Beagle 4.1 (Yan et al. 2017) was higher than our imputation accuracy of 0.71 for imputation from 70k to WGS (Table 1).

Conclusions

Imputation from commercial SNP chips to WGS in pigs is possible, however imputation accuracy is limited by the current size of the available reference population. Imputation from (genotyped or accurately imputed) 660k is preferred to imputation from 70k directly to WGS. Imputation of variants with low MAF is problematic, hence results for these variants should be treated with caution. With regard to imputation programs, Minimac3 seemed to perform best, but with a closer look at the details, Beagle 4.1 outperformed Minimac3. FImpute performed less well, but was not used to its full potential here as no pedigree information was provided, and the studied population did not have the same properties as in practical settings that require imputation. With respect to the latter aspect, it is important to realise that in practise imputation accuracy can be better than shown here, especially the two-step procedure and FImpute results.
Table 1. Mean imputation accuracy ($r$) of SSC7 from 70k (3356 SNP) or 660k (29,249) genotypes to WGS (1,664,543 variants) using different imputation programs.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Program</th>
<th>Nr variants</th>
<th>$r$</th>
<th>Nr overlapping variants</th>
<th>$r$ (overlapping variants)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70k-WGS</td>
<td>Beagle 4.1 ML$^1$</td>
<td>1,046,343</td>
<td>0.708</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>70k-660k</td>
<td>Beagle 4.1 ML</td>
<td>26,763</td>
<td>0.839</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>70k-660k-WGS</td>
<td>Beagle 4.1 ML</td>
<td>1,046,343</td>
<td>0.688</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>660k-WGS</td>
<td>Beagle 4.1 ML</td>
<td>1,019,754</td>
<td>0.810</td>
<td>859,296</td>
<td>0.909</td>
</tr>
<tr>
<td>660k-WGS</td>
<td>Fimpute ML</td>
<td>1,022,332</td>
<td>0.742</td>
<td>859,296</td>
<td>0.839</td>
</tr>
<tr>
<td>660k-WGS</td>
<td>Minimac3 ML</td>
<td>859,469</td>
<td>0.889</td>
<td>859,296</td>
<td>0.889</td>
</tr>
<tr>
<td>660k-WGS</td>
<td>Beagle 4.1 DS$^2$</td>
<td>1,020,588</td>
<td>0.821</td>
<td>859,296</td>
<td>0.917</td>
</tr>
<tr>
<td>660k-WGS</td>
<td>Minimac3 DS</td>
<td>1,020,535</td>
<td>0.788</td>
<td>859,296</td>
<td>0.910</td>
</tr>
</tbody>
</table>

$^1$ ML=most likely genotype (i.e., 0, 1 or 2)
$^2$ DS=allele dosage (any value between 0 and 2)

Figure 1. Imputation accuracy as a function of MAF. Solid line is Minimac3 ML, dashed line is Beagle 4.1 DS, dotted line is Beagle 4.1 ML, dot-dashed line is Minimac3 DS, and long dashed line is FImpute ML.
Figure 2. Smoothed scatterplot of imputation accuracy for Beagle 4.1 ML along SSC7 (A), in the region from 24 to 27Mbp (B), and in the region from 83 to 85Mbp (C). Red line is the LOESS smoothed line. Green dots at bottom indicate the position of SNP from the 660k SNP chip.
List of References


