Significance thresholds and genomic control for a GWAS using SNP panel or imputed whole genome sequence data of a pig population

S. van den Berg1, J. Vandenplas2, F.A. van Eeuwijk2, M.S Lopes3, R. F. Veerkamp1

1 Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, the Netherlands
2 Wageningen University & Research Biometris, 6700 AH Wageningen, the Netherlands
3 Topigs Norsvin Research Center, 6640 AA, Beuningen, the Netherlands
sanne.vandenberg@wur.nl (Corresponding Author)

Summary

Determining the appropriate significance threshold for a genome wide association study, usually based on the −log10(p-value), is important to minimize the chance of finding false positive associated SNP. So far, there is little consensus in the field of pig breeding on the appropriate threshold and the best approach to correct for false positives when using sequence information. Therefore this study aimed to compare different adjustments of significance thresholds and genomic control in order to minimize the occurrence of false positives for QTL detection in a pig breeding population using medium and high density SNP panels or imputed whole genome sequence data. From our results, we recommend to perform genomic control of each chromosome independently and to use a significance threshold based on the Bonferroni correction with the total number of SNPs since it takes into account the increase in number of SNP better than accounting for independent chromosomal segments, even though the underlying assumption seems to be too conservative.

Keywords: GWAS, false positives, Bonferroni correction, permutation testing, genomic control

Introduction

Genome-wide association studies (GWAS) aim to associate single nucleotide polymorphism (SNP) with a trait of interest in order to get a better understanding of the genetic architecture and to improve genomic prediction (VanRaden et al. 2017). The number of false positive SNP is however expected to increase when whole genome sequence (WGS) is used instead of a lower density SNP chip. Therefore it is important to determine an appropriate significance threshold, usually based on the −log10(p-value), to minimize the occurrence of false positives.

In human genetic studies a significance threshold of −log10(p-value) > 5 (Welter et al. 2013) is commonly accepted. However this significance threshold is unlikely to be appropriate for data from a pig breeding program because strong family relationships exists, causing long range linkage disequilibrium (LD) between SNPs. Therefore the significance threshold can be less conservative. So far there is no consensus in literature about the appropriate significance threshold in a pig breeding population, e.g. the −log10(p-value) varied from 3.3 to 6 using either no multiple testing correction, a Bonferroni correction or genomic control (e.g. van Son et al. 2017; Hao et al. 2017; Le et al. 2017; Do et al. 2013;

To minimize the discovery of false positives, permutation testing is a gold standard method to derive a empirical significance threshold that allows for the statistical dependence between SNPs. It involves the break down of the association between phenotypes and genotypes by randomly shuffling phenotypes (Churchill & Doerge 1994). However, permutation testing is computationally expensive. An alternative approach to permutation testing to determine the significance threshold is the Bonferroni correction that adjusts significance thresholds by the number of independent tests. A conservative way is assuming that the number of independent tests is equal to the number of SNPs tested as number of independent tests. However, this approach ignores LD between SNPs, which is especially present when strong family relationships exist. Recognizing the presence of LD between SNPs, Duggal et al. (2008) suggested to assume that the number of independent tests is equal to the number of independent chromosome fragments, since these segments describe the regions on the genome that explain unique genetic variation.

In addition to determining the appropriate significance threshold, genomic control can help to minimize the discovery of finding false positives. Genomic control is quality control of the p-values and is perfomed based on the genomic inflation factor, that expresses the deviation of the observed p-values compared to the expected p-values. Inflated p-values indicate a high chance of finding false positives. Since different LD patterns across chromosomes may exist, we would expect that genomic inflation factors differ across chromosomes.

Therefore, the objective was to compare different adjustments of significance thresholds and genomic control across chromosomes in order to minimize the chance of finding false positives for GWAS in a pig breeding population using SNP panels or iWGS data.

**Material and Methods**

**Data**

Dataset for a Large White (LW) line was provided by Topigs-Norsvin. The dataset included pre-corrected phenotypes for number of teats of 4964 Large White (LW-line) pigs (see Lopes et al. 2017 for more details on phenotypes and pre-correction), medium density genotypes (34,588 SNPs) and high density genotypes (491,169 SNPs). High density genotypes were imputed to WGS with Beagle 4.0 (Browning & Browning 2009), using a multi-line reference population of 168 animals of which 32 individuals originate from the LW line. After imputation and quality control, 10,212,687 SNPs that had a Beagle imputation accuracy (R²) > 0.6 were considered for iWGS. The average R² was 0.93.

**Single-SNP genome-wide association study**

A single-SNP GWAS was performed for medium density, high density, and iWGS, applying a mixed linear association model with a leave one chromosome out (LOCO) approach as implemented in GCTA version 1.25.2 (Yang et al. 2011; Yang et al. 2014). The model was as follows:

$$ y = \mu + b + x + \mathbf{u} + \mathbf{e} $$

where $y$ is a vector of the phenotypes, $\mu$ is the mean, $b$ is the fixed effect of the SNP tested for association, $x$ is a vector of the SNP genotypes coded as 0, 1, or 2, $\mathbf{u}$ is a vector of random polygenetic effect, and $\mathbf{e}$ is a vector of residuals. The residuals are distributed following a
normal distribution \(N(0, \sigma^2)\) with \(\sigma^2\) being the residual variance. The vector for the random polygenetic effect \(u\) followed a normal distribution \(N(0, G \sigma^2)\) where \(G\) is the genomic relationship matrix (Yang et al. 2010) for which the chromosome of the SNP tested is ignored, and \(\sigma^2\) is the genetic variance.

**Permutation testing**

The procedure for permutation testing followed Churchill & Doerge (1994). The phenotypes were randomly shuffled and subsequently used for a single SNP GWAS analysis. A total of 1000 replicates were performed, and the maximum \(-\log_{10}(p\text{-value})\) of each replicate was recorded. The \(-\log_{10}(p\text{-value})\) significance threshold was defined as the 95th percentile of the ordered recorded values. The permutation test was performed for chromosomes 4, 7 and 10 using either medium density genotypes, high density genotypes, or iWGS. Only three chromosomes were tested because of computational costs.

**Bonferroni correction**

The Bonferroni correction divides the probability of having at least one false-positive result when the null hypothesis \(H_0\) is true \((\alpha)\) by the number of independent tests. In this study, \(\alpha\) is assumed to be 0.05. The number of independent tests was either the total number of SNPs or the number of independent chromosome fragments. The number of independent chromosome fragments \((Me)\) was calculated as follows:

\[
(2)
\]

where \(G\) is the genomic relationship matrix considering all SNPs and \(A\) is the pedigree relationship matrix. The genomic relationship matrix was computed following the first method of VanRaden (2008).

**Genomic inflation factor**

The chi-square tests statistics, needed for the computation of the genomic inflation factors, were calculated from the p-values assuming 1 degree of freedom. The genomic inflation factor was defined as the median of the observed chi-squared test statistics divided by the expected median of the corresponding chi-squared distribution, and was computed for each chromosome and for the whole genome for the different densities.

**Results**

**Significance thresholds**

The significance thresholds, expressed as \(-\log_{10}(p\text{-value})\), found with permutation testing increased from 4.1 to 5.5 when the marker density increased from medium to iWGS. The significance thresholds were constant between chromosomes (Table 1.). The significance thresholds derived from the Bonferroni correction using the total number of SNPs increased from 5.8 to 8.3 with increasing marker density (Table 2.), but remained similar across the different SNP densities, that is about 3.7, when the number of independent chromosome fragments were used for the Bonferroni correction.

**Genomic inflation factor**
At a whole genome level, the genomic inflation factor is higher than 2 and is about the same for the medium and high densities and iWGS (Figure 1.). Similarly, at the chromosome level, the genomic inflation factors remained constant across the different densities. However, the genomic inflation factors varied between chromosomes, from 0.979 to 4.181 (Figure 1.). For permutation testing, average genomic inflation factors over 1000 permutations were around 1 and remained constant across chromosomes and densities (Table 1.).

Discussion

Significance thresholds

The significance threshold established with permutation testing and with the Bonferroni correction based on the total number of SNPs increased with increasing SNP density. This suggests that the Bonferroni correction might be an appropriate and less computational intensive alternative to the permutation testing. However this Bonferroni correction ignores LD between SNPs whereas the permutation testing does account for it. Also, it resulted in a far too high threshold (e.g. 5.4 with permutation testing and 8.4 with the Bonferroni correction for iWGS). Therefore, the significance thresholds based on this Bonferroni correction is too conservative, leading to undetected associated SNPs.

The Bonferroni correction with the number of independent chromosome fragments does account for LD between SNPs. Compared to the significance thresholds established with permutation testing, the significance thresholds based on this Bonferroni correction is underestimated for all densities (e.g. 5.4 with permutation testing and 3.4 with the Bonferroni correction for iWGS), as also shown by Dudbridge & Gusnanto (2008). They argued that their method to estimate the number of independent chromosome fragments might not be right. In this study the number of independent chromosome fragments was calculated following the formula proposed by Goddard et al. (2011). This formula assumes uniformity of LD patterns across and within chromosomes. This assumption might be invalid since LD patterns can differ across and within chromosome (Veroneze et al. 2013), which can lead to underestimated values, and an increase of the occurrence of false positives.

Overall, the best option to account for multiple testing is to use the total number of SNPs in the Bonferroni correction, even though this threshold is too conservative.

Genomic inflation factors

In this study we found genomic inflation factors that were above 2 for lower density SNP genotypes and iWGS. High inflation factors can be caused by strong population stratification, unexplained relatedness and systematic bias (Devlin & Roeder 1999; Zheng et al. 2006; Reich & Goldstein 2001). Hinrichs et al. (2009) have reported that the genomic inflation factors can be controlled by accounting for population stratification. They also reported that strong LD and strong association between SNPs and a trait of interest can cause problems with controlling inflation factors. In this study high inflation factors might be caused by the strong association between SNPs and the trait of interest, since the inflation factors decreased to one with permutation testing which removes the association between SNPs and the phenotype. A solution to remove strong association between SNPs and phenotype might be to fit the large QTL as a fixed effect in the model.

Genomic inflation factors varied largely across the chromosomes, which might
indicate that genomic control should be applied per chromosome. Indeed, using the genomic inflation factor of the whole genome can potentially bias association tests conservatively in some regions and freely in other regions. Based on these results, it is recommended to apply genomic control per chromosome.

**Conclusion**

The objective was to compare different adjustments of significance thresholds and genomic control for GWAS in order to minimize the chance of finding false positives in a pig breeding population using medium and high density SNP panels or iWGS data. From our results, we recommend to perform genomic control per chromosome and to use a significance threshold based on the Bonferroni correction with the total number of SNPs since it followed the same trend as the significance threshold found with permutation testing, even though the underlying assumption seems to be too conservative.

**Acknowledgements**

The authors want to acknowledge Topigs Norsvin for providing the data. The authors also want to thank the Netherlands Organisation of Scientific Research (NWO) and the Breed4Food consortium partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin for their financial support.

**Tables**

*Table 1. Significance thresholds and genomic inflation factors from permutation testing of chromosomes 4, 7 and 10 for medium and high densities and iWGS*

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Density</th>
<th>p-value threshold&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Genomic inflation factor&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Medium</td>
<td>4.178</td>
<td>0.997 (0.198)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>4.927</td>
<td>1.004 (0.204)</td>
</tr>
<tr>
<td></td>
<td>iWGS</td>
<td>5.469</td>
<td>0.991 (0.196)</td>
</tr>
<tr>
<td>7</td>
<td>Medium</td>
<td>4.232</td>
<td>1.003 (0.207)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>4.922</td>
<td>1.003 (0.201)</td>
</tr>
<tr>
<td></td>
<td>iWGS</td>
<td>5.449</td>
<td>1.005 (0.210)</td>
</tr>
<tr>
<td>10</td>
<td>Medium</td>
<td>4.100</td>
<td>1.003 (0.194)</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>4.743</td>
<td>0.988 (0.188)</td>
</tr>
<tr>
<td></td>
<td>iWGS</td>
<td>5.426</td>
<td>0.988 (0.185)</td>
</tr>
</tbody>
</table>

<sup>1</sup> p-value thresholds are expressed as –log10(p-values)

<sup>2</sup> Averages and SD within brackets over 1000 permutations

*Table 2. Significance thresholds of a Bonferroni correction using the total number of SNPs or the number of independent chromosome fragments for medium and high densities and iWGS*

<table>
<thead>
<tr>
<th>Bonferroni total&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Bonferroni_ M&lt;sub&gt;c&lt;/sub&gt;&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td># SNPs</td>
<td>Significance</td>
</tr>
</tbody>
</table>

---

**Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.994**
<table>
<thead>
<tr>
<th></th>
<th>threshold</th>
<th>threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium density</td>
<td>34,588</td>
<td>5.840</td>
</tr>
<tr>
<td></td>
<td></td>
<td>276.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.743</td>
</tr>
<tr>
<td>High density</td>
<td>491,169</td>
<td>6.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>255.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.708</td>
</tr>
<tr>
<td>iWGS</td>
<td>10.2 M</td>
<td>8.310</td>
</tr>
<tr>
<td></td>
<td></td>
<td>275.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.741</td>
</tr>
</tbody>
</table>

1. Bonferroni total = 0.05/total number of SNPs
2. Bonferroni $\text{Me} = 0.05 / \text{Me}$
3. $\text{Me}$ is the number of independent chromosome fragment calculated with the formula proposed by Goddard et al (Goddard et al. 2011)
4. Significance thresholds are expressed as $-\log_{10}(p$-values)

**Figures**

*Figure 1. Inflation factors per chromosome and the total genome found with medium density, high density and iWGS.*

**References**


threshold to control the family-wide type I error in genome-wide association studies. BMC Genomics. 9: 516.


