Mapping Resolution in Single and Multiple Porcine F2 Populations using Genome Sequence Marker Panels

Jörn Bennewitz[1] and Robin Wellmann[1].
[1]Institute of Animal Husbandry and Breeding, University Hohenheim, Germany

ABSTRACT: Porcine F2 crosses derived from distantly or closely 'related' founder breeds have frequently been used for QTL mapping. Increasing marker density up to a maximum level can be done by re-sequencing and imputation techniques. This study investigated by means of simulations the mapping resolution and LD structure around causal genes of several F2 crosses with maximum marker density. It is shown that the mapping resolution is high (low) in F2 cross from closely (distantly) 'related' founder breeds. Poling data from several crosses improves mapping resolution substantially. In addition, it is shown that for causal genes segregating in a particular founder breed, the number of markers being in high LD is smaller in pooled F2 crosses than in the founder breed itself. This high mapping resolution makes pooled F2 crosses with maximum marker density suitable for identification of segregating causal genes.

Keywords: pooled F2 crosses; association mapping; mapping resolution

Introduction

During the last two decades in pig breeding many QTL mapping experiments were conducted (Rothschild et al. (2007)). The experimental design was frequently an F2-cross established from two genetically divergent and outbred founder pig breeds. The founder breeds were frequently chosen from the Asian type and from the European type of breeds. Phylogenetic analysis of whole genome sequence data revealed distinct lineages of these two types of breeds (Frantz et al. (2013). But also F2-crosses within European type of breeds were established (e. g. Boysen et al. (2010)). In general, individuals were mainly genotyped with microsatellite markers and QTL mapping relied on linkage between marker and QTL. The power to map QTL precisely is limited, because of the limited number of individuals included in a typical F2 cross, limited number of useable meiosis, and the use of low density marker panels together with the use of linkage information only. One way to increase the power is to pool data from different F2 crosses and to jointly analyse the data (Rückert and Bennewitz (2010)).

Association mapping relies on linkage disequilibrium (LD) between marker and QTL and, in contrast to linkage analysis, utilizes also historical meiosis (e. g. Goddard and Hayes (2009)). A high marker density is necessary for genome wide association studies (GWAS). Marker density is maximised if the sequence of the individuals is know. In this case the aim is to find the causal mutations in the pool of all mutations and to separate them from the other markers which are in LD with the mutation. Large scale re-sequencing is still unaffordable, but sequence information can be imputed using SNP chip data. In procine F2 crosses, the Illumina PorcineSNP60 BeadChip (Ramos et al. (2009)) with around 62K SNPs can be used to accurately impute sequence data from founder individuals in F1 and subsequently in the F2 generation using mainly pedigree information. Hence, sequence data of F2 individuals can be generated by only re-sequencing the founder individuals and chip genotyping the F1 and F2 generation, which is affordable in many situations. Many crosses were established from distantly 'related' founder breeds (e. g. from an Asian and a European type breed). Long LD blocks are observed in these crosses. Pooling of several F2 crosses might reduce the length of LD blocks and hence increases mapping resolution in these crosses. In contrast, in F2-crosses established from closely 'related' founder breeds (e. g. from two European type breeds), the blocks with a high LD are short and might even be shorter compared to outbred populations (Toosi et al. (2010)). Hence, it still might be worthwhile to continue using well established F2 outbred crosses also for GWAS with maximum marker density, especially if founder breeds are 'related' or data from several F2 crosses can be pooled and analysed jointly or both.

Evaluating the power to map genes by means of stochastic simulations is desirable and was done for a single porcine F2 cross situation by Ledur et al. (2009). It requires making numerous assumptions and decisions about the use of the mapping procedure (e. g. single marker vs. multi-marker approaches), threshold levels and determination of thresholds. However, as stated above, the power is a function of the marker density and LD-structure around the causal mutations in the final mapping population.

The aim of this study was to analyze the mapping resolution of single and joint F2 outbred crosses by means of stochastic simulations and in situations where marker density is maximised. We analyzed the number of segregating SNPs and causal genes, the pattern of LD within and across pooled F2 crosses, and investigated the LD structure in the vicinity of causal genes. In addition we investigated whether focusing solely on genes that are segregating in a particular founder breed of interest is a suitable strategy for maximising mapping resolution for these causal genes.

Materials and Methods

We simulate the three F2 crosses that were established at the University in Hohenheim (Germany, Rückert and Bennewitz (2010)), which were a Meishan (Pop M) x Pietrain (Pop P) cross (MxP), a European wild boar (Pop W) x Pietrain cross (WxP), and a WxM cross. In these crosses the founder breeds are distantly 'related'. In addition we simulated the F2 cross established at the University in Kiel (Germany, Boysen et al. (2010)), which where Large White (Pop LW) and Landrace (Pop L) crossbred individuals were mated to P individuals and developed towards an F2 cross (Px(LxLW)). We chose these four F2 crosses as a reference because they were

generated from distantly and closely 'related' founder breeds, respectively, and merging them into a single large data set for mapping purposes is an issue. The traits collected on the Kiel and Hohenheim F2 individuals are important for sire breeds and P is used heavily as a sire breed in crossbreeding schemes. It was selected for these traits and is the most important founder breed from an economic point of view. Hence, special attention will be put on alleles segregating within P.

We started with the simulation of the phylogeny of the 5 populations, M, P, W, LW, and L. The following simulation protocol is based on what is known about the phylogeny of pig breeds (e. g. Frantz et al. (2013)). The ancestral population with N_e = 2000 was maintained for 8000 generations to reach mutation-drift equilibrium. Thereafter it was split 2000 generations before present (GBP) into three subpopulations with N_e = 600. These three subpopulations represent W, M, and the European domestic pigs (P, L, LW). By assuming a generation interval of 2.5 years the most recent common ancestor of these three simulated populations lived 5000 YBP. The subpopulation representing the three European domestic pigs was split 100 GBP (corresponding to 250 YBP) into 3 additional populations representing LW, L, and P with N_e = 400. The N_e of M also decreased to N_e = 400 at this time. The N_e of all domestic populations (i.e. all populations except W) was reduced 20 GBP (50 YBP) to N_e = 200.

These populations were used to produce the four F2 crosses mentioned above. For producing Px(LxLW), 24 females were obtained by crossing populations L and LW. These 24 females were mated with 3 boars from P to produce the F1 consisting of 15 males and 120 females. The individuals from the F1 were mated to produce the F2 consisting of 1200 individuals.

Each of the crosses WxP, MxP, and WxM consisted of 400 F2 individuals that were obtained from 40 F1 females and 5 F1 males. The F1 individuals were obtained from one purebred male and 8 purebred females. For WxP the male was from W and the females were from P. For WxM the male was from W and the females were from Pop M. Finally, for MxP the male was from M and the females were from P. Note that the number of founders of cross Px(LxLW) is equal to the total number of founders in crosses WxP, MxP, and WxM. Moreover, the number of F2 individuals in cross Px(LxLW) is equal to the total number of F2 individuals in crosses WxP, MxP, and WxM. These simulated pedigree structures reflect the structure of the real existing crosses (Rückert and Bennewitz (2010); Boysen et al. (2010)).

The individuals had 1 chromosome with a length of 1 Morgan. The expected number of new mutations per individual was nMut = 1.5. Each new mutation was a causal gene with probability pQTL = 0.1. The gene effects of new mutant genes were normally distributed and independent. The traits were purely additive. Errors were added to the genotypic values to obtain traits with average heritability 0.5. The number of simulated genes was rather large in order to get smaller prediction errors of the average number of markers in a high LD with a single causal gene. Pop P was selected by truncation selection within males in the last 20 generations for one trait. Consequently, the total population size was larger than N_e for this population in the last 20 generations. The r^2-value between two SNP was computed as the squared correlation between the haplotype alleles. The alleles were coded as 0 and 1. An SNP remained in the data set if it had MAF > 0.01 in the founder animals (across all populations). It was considered as segregating in the respective population / cross if both alleles occurred within the population / cross. Various statistics characterising the number of SNPs and genes as well as the mapping resolution and LD structure around causal mutations were calculated from the simulated crosses.

Results and Discussion

Table 1 shows the number of SNPs segregating in the respective populations and the genetic connections between the different populations. Note that the number of causal genes segregating in the population was around 10% of the number of SNPs. Most SNPs are segregating in W, which was expected because this population had the largest historic N_e. The number of SNPs segregating in M is smaller than the number of segregating SNPs in LW, W, and M. This is due to ascertainment bias in the selection of segregating SNPs from the simulated sequence data (i.e. MAF > 0.01 in the founder animals). M had more private alleles, as shown in the second column. This column contains the probability that an SNP which is segregating in the respective population is also segregating in another population. The third column contains the conditional probability that an SNP is segregating in the respective population, given that it is segregating in P. As stated above, P is the breed of interest from a breeding perspective. SNPs segregating in this population are very likely to segregate also in populations L and LW. The probability that it is segregating in W or M is small but not negligible.

Table 1. Average number of SNPs in the simulated founder populations and genetic connections between simulated founder populations

Simulated population	SNPs	Prob. seg. elsewhere	Prob. seg. cond. that it is seg. in P
Wild boar, W	8279	0.162	0.069
Meishan, M	5670	0.209	0.060
Large White, LW	6699	0.915	0.801
Landrace, L	6774	0.920	0.818
Piétrain, P	6724	0.904	1.000

Average number segregating SNPs (first column), probability that an SNP segregates in the population and simultaneously in another population (second column), and conditional probability that a SNP segregates in the population given that it segregates in Piétrain population (third column).

Table 2 shows the number of segregating SNPs in the first column. Compared to the founder breeds, this number is substantially higher. In the next columns the average number of neutral SNPs that are in high LD (r^2 > 0.95) with a single causal gene, averaged over the simulated causal genes with MAF > 0.05 are shown. In column 2 and

3 all SNPs and genes segregating in the respective set of individuals (founder breed P, single crosses or pooled crosses) are involved. In the columns 'Distant LD SNPs', only the SNPs having a distance of at least 0.01 Morgan from the causal gene are counted. These columns show that the number of LD SNP and of Distant LD SNPs is low in P. The number of LD SNPs is however, even lower in the cross Px(LxLW), because the founder breeds of this cross are related. This demonstrates the high general mapping resolution in this cross. This is in agreement with Toosi et al. (2010). For the remaining single crosses these figures are higher and thus mapping resolution is substantially smaller.

In addition, the last two columns of Table 2 contain SNPs and genes only if they are segregating also in the important breed P. In a breeding program for P only these two columns are of interest. Considering the joined F2 crosses instead of the purebred population P decreases the number of neutral markers being in high LD by around two third. Hence mapping resolution is highest for these important genes in the pooled F2 crosses.

Conclusion

We showed that pooling F2 populations with maximum marker density is a suitable strategy for obtaining low short range LD between alleles that are segregating in a founder breed (investigated in this study for the important breed P). The results imply that the pooled F2 cross is even more suitable for identification of the causal mutation than the parental population. This conclusion, however, holds only for genes segregating in a founder breed. Using pooled F2 crosses for identification of causal genes that are fixed in the founder breeds still suffers from the small number of recombinations by creating the F2 crosses.

For mapping genes not only the LD structure is important but also the genetic architecture of the trait. The effect of a gene may depend on the genetic background, e.g. on the breeds that are used for creating the F2 crosses. Pooling the crosses may require a regression model that allows a marker to have different but correlated effects in the different crosses. We are in progress of applying GWAS to these simulated data sets and to develop models that cope with these complex population structures.

Authors Contribution

Both authors conceived the study and wrote the paper. RW did the statistical analysis.

Literature Cited

Boysen, T. J., Tetens, J., and Thaller, G. (2010). J. Anim. Sci. 88:3167-3172

Frantz, L. A. F., Schraiber, J. G., Madsen, O. et al., Genome Biology, 14(9): R107

Goddard, M. E., and Hayes, B. J. (2009). Nat. Rev. Genet. 10:381-391.

Ledur, M., Navarro, N., and Pérez-Enciso, M. (2009). Heredity 105:173-182

Ramos, A. M., Crooijmans, R. P. M. A., Affara, N. A. et al. PLoS One, 4:8, e6524.

Rothschild, M. F., Hu, Z.L., and Jiang Z. (2007). Int. J. Bio. Sci. 3:192-197

Rückert, C., and Bennewitz, J. (2010). Genet. Sel. Evol. 42:40.

Toosi, A., Fernando, R. L., and Dekkers, J. C. M. (2010). J. Anim. Sci. 88:32-46

Table 2. Average number of SNPs and SNPs in high LD with a single causal gene in the simulated populations

Simulated population and cross	SNPs	Total LD SNPs	Distant LD SNPs	Total LD SNPs seg. in P	Distant LD SNPs seg. in P
P	6724	4.21	0.02	4.21	0.02
Px(LxLW)	7619	3.16	0.09	3.14	0.06
WxP	11654	64.57	18.71	6.97	0.83
MxP	11502	61.41	16.23	8.15	0.90
WxM	10600	65.95	21.72	11.52	3.66
WxP & MxP & WxM	17388	32.35	11.95	3.42	0.17
Four crosses pooled	19178	38.01	20.09	1.36	0.00

Average number segregating SNPs (first column), average number of SNPs that are in high LD ($r^2 > 0.95$) with a single causal gene (second and third columns) and average number of LD SNPs that segregate in the simulated Piétrain (fourth and fifth columns). The columns *Distant LD SNPs* refers to average number of SNPs in high LD with a single causal gene and having a distance > 0.01 Morgan from the causal gene.

[1] P Piétrain, L Landrace, LW Large White, W European wild boar, M Meishan.