SNP Effects Depend on Genetic and Environmental Context

J. W. M. Bastiaansen[*], **H. Bovenhuis**[*], **M. S. Lopes**[*†], **F. F. Silva**[‡], **H-J. Megens**[*], **M. P. L. Calus**[§]
[*]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands; [†]TOPIGS Research Center IPG, Beuningen, the Netherlands; [‡]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Brazil; [§]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands.

ABSTRACT: Effects that are estimated for SNP markers depend on LD with the QTL, and interactions of the QTL with other genetic and environmental factors. These factors are often mentioned but rarely studied. Breeding for crossbred performance both brings the need and supplies data for studying these interactions. SNPs with different effects on litter size in pigs between low and high production environments were identified from a genomic reaction norm model. Clustering of these SNPs lead to candidate genes related to bacterial defense that are expressed in reproductive tracts and regulated by the estrous cycle. To study interaction of SNPs with genetic background, a method to determine breed origin of alleles in crossbreds was implemented using long range phasing with AlphaPhase software. With more genotypes and phenotypes on crossbreds, estimation of interactions with genetic background and the environment will become feasible.
Keywords: SNP effects; G×E Interaction; Epistasis

Introduction

Marker-trait associations have been studied for a long time (e.g. Sax (1923); Andersson et al. (1994)) and are at the center of many studies currently performed in animal genetics. Both the research areas of association analyses and genomic selection rely on association between markers, most often SNPs, and phenotypes to either discover the genetic architecture of traits or predict future performance of animals. In this paper we used the term SNP as it is currently the most frequently used type of marker.

A number of factors have an impact on the effect that can be measured for a SNP and they are often mentioned when results from one study do not transfer to other studies. For example, when peaks that are found in a genome wide association study (GWAS) in breed A are not replicated by a GWAS in breed B the reasons mentioned may include one or more of the following three arguments. First, the associated SNP(s) in breed A are not in Linkage Disequilibrium (LD) with the QTL in breed B. Second, the genetic background in breed B is 'different'. This could mean that the functional variation that underlies the QTL is not segregating in breed B or epistatic interactions in breed B are different due to other genes that modify the effect of the QTL in breed B. Third, the phenotype being measured in breed B is not the same as in breed A. This could simply

be because the definition or measurement method of the trait is not the same (Barendse (2011)), but there could also be environmental differences experienced by animals from breed B that cause different QTL to be important for the trait when measured in breed B, i.e. genotype by environment interaction (G×E) is present. Presence of G×E is particularly evident when one of the breeds is kept in, say, a breeding facility and the other breed or cross is kept under commercial production conditions, possibly in a different country and/or climate.

Besides these three reasons, that are of interest to be unraveled by geneticists, there are other, less exciting reasons for lack of replication such as lack of power in the analysis of breed B, or spurious results in breed A (for more reasons, see Chanock et al. (2007)). The results in breed A could be spurious because of 'noise' such as hidden structure in the data that is not accounted for in the analysis. Experimental design and statistical analysis will normally attempt to remove this noise. However, one specific form of this 'noise' are the associations due to family relationships. While these family relationships may lead to false positives in association analyses, they are in fact of use in the context of genomic prediction when selection candidates are related to the reference population.

Knowledge about the impact of LD, and especially of epistasis and G×E on the effects of QTL and the estimates of SNP effects is limited. Considering that these factors are so often mentioned in the discussion of GWAS and genomic prediction studies, more studies into the extend and importance of these effects are warranted.

Our first objective was to introduce and discuss the impact on SNP effects from the three factors, LD, interaction of genotype with the environment (G×E) and with genetic background (epistasis). Second we report results from two studies that aimed to discover the effects of G×E and epistasis on SNP effects. The first study applies functional clustering to discover genes affecting the change in SNP effects across different environments. The second study applies long range phasing to trace the breed origin of alleles in crossbred animals.

Variation in LD. The reason for finding an association of a SNP with the phenotype is the (nearby) presence of genetic variation that has a functional effect on the phenotype, e.g. one or more QTL in LD with the SNP.

Differences in LD can obviously lead to differences in the effects that are estimated for a specific SNP in different populations, but should by itself not lead to lack of replication of QTL results. When analyses are performed separately for each population with a marker density that is high enough for some SNP to be in LD with the QTL, associations should replicate. These associations do not need to be with the same SNP in each population. Differences in LD between breeds can actually be used to move towards identification of Quantitative Trait Nucleotides (QTN), provided that the effect of such QTN is present in the other breeds, without modification due to interactions with genetic background and environment. Ciobanu et al. (2001) described that estimated haplotype effects at the PRKAG3 locus were more consistent across populations and resolved the effects of PRKAG3 more clearly than the individual SNPs. Individual SNP effects were more variable between populations, due to differences in haplotype frequencies.

When interest is in genomic prediction across populations, the differences in LD are an important issue (de Roos et al. (2009)). Persistence of LD between populations is a requirement when the aim is to predict breeding values. The persistence of LD will be improved with increasing density of SNPs but evidence that this leads to higher prediction accuracies is thus far limited.

Genotype by environment interaction. G×E usually refers to the changes in the ability of animals to perform under different environments. In pig breeding this is often referred to as 'robustness' (Knap (2005)), meaning the lack of G×E interactions. In plant breeding G×E interaction leads to the selection of specific lines for specific environments but in animal breeding this is largely avoided.

While extensive work has been done on the effect of environment and G×E interactions at the animal level, a limited number of studies investigated the impact of SNPs on environmental sensitivity of animals (Lillehammer et al. (2009), Streit et al. (2013)) or changes in SNP effects due to G×E (Lillehammer et al. (2009)). A recent study in dairy cattle showed that mutations in the DGAT and SCD genes had a different effect on milk fatty acid composition during winter and summer conditions (Duchemin et al. (2013)). Even though the G×E effects of DGAT and SCD were relatively small and genotypes were not re-ranked, it is sensible to expect that the effect of a SNP changes with the environment. Simply put, when the mean phenotype increases from one environment to the next then the size of the individual QTL effects must also increase if everything else, such as the allele frequencies, interactions between QTL, and the level of heritability, stay the same.

Recently a genomic reaction norms approach was described that modeled breeding values over a range of environmental levels and resulted in an estimate of the slope and intercept of the animal breeding values (Silva et al. (2014)). The method was applied to records of total number born (TNB) from daughters of genotyped sires that were used for inseminations across many farms and countries. A two-step reaction norm approach (Calus et al. (2002) was applied where the first step resulted in corrected phenotypes of the sows and in estimates of the herd, year, season (HYS) effects. The second step estimated the breeding values for slope and intercept of each sire with a random regression model, using the corrected sow TNB phenotypes and the HYS estimates from the first step as input. The second step was carried out both with the pedigree relationship matrix **A** and the genotypes relationship matrix **G**. Accuracies of the sire's estimated breeding values were improved both by the use of the G×E model over the standard model, without the HYS, and by the use of the **G** over the **A** matrix (Silva et al. (2014)). Here we further investigated the individual SNP effects. Specifically, we investigated whether the SNPs that show the largest change in effect over environments lead us to genes that influence the interaction with environment.

Interaction with genetic background. In a recent review, Mackay (2014) showed that epistatic interactions between loci must exist, and that knowledge of these interactions will, among others, improve our ability to predict long term response to selection as well as explain heterosis and inbreeding depression. In model species, epistasis has been shown to be pervasive (Mackay (2014)). Within a breed the epistatic variance is due to alleles segregating at, at least, two loci. Differences in SNP effects between breeds that are due to epistasis, however, do not require that the other loci that interact with the SNP segregate within the breeds. Alternate alleles between breeds at the other loci is sufficient. The interaction effects may be difficult to detect but the breed specific effect of a SNP, may still be considerable in size.

In pigs, Su et al. (2012) showed considerable non-additive contributions to variation in daily gain. Besides a dominance variance accounting for 5.6%, additive by additive epistatic variance was shown to account for 9.5% of total phenotypic variance in purebred Duroc pigs. These levels of epistatic variance, in combination with known differences in allele frequencies between pig breeds (Wilkinson et al. (2013)) predict considerable differences of SNP effects between breeds due to interaction with the genetic background.

The use of crossbred production animals is both a reason and opportunity to study interaction effects of SNPs with the genetic background. The effect of the same allele could, in principle, be compared when present in the pure breed and in the crossbreds. Because SNPs often segregate in many of the pig breeds, a prerequisite for such an analysis is that the purebred origin of alleles in crossbred can be determined. Here we describe an approach and results for assigning breed origin to alleles in crossbreds.

Materials and Methods

Although presented here as separate research questions with separate experiments, the interactions of SNPs with the environment and with the genetic background are often confounded. In our experiments with

crossbred data the two effects will simultaneously be present. The genetic background can more easily be standardized when estimating interactions with the environment than vice-versa. When studying differences between SNP effects in pure and crossbreds for better prediction of performance we may be less worried about the genetic or environmental origin of the interaction effect.

SNP by environment interaction. The study by Silva et al. (2014) focused on the accuracy of genomic prediction when fitting a genomic reaction norm to phenotypes of TNB. Phenotypes were obtained in different environments (e.g. HYS levels) ranging from 10 to 22 TNB. From that study we obtained estimates of slope and intercept for each SNP. When the effect of SNP depends strongly on the environment we should get a large effect for slope. The top 100 SNPs for absolute value of slope were selected and all genes identified within $\pm$ 500 kilobases (Kb) on the Sscrofa10.2 assembly of the reference genome (Groenen et al. (2012)) were retrieved using the BioMart interface on the Ensembl Genes 75 database (Flicek et al. (2014)). The gene names were subsequently used in a functional annotation clustering implemented in DAVID Bioinformatics Resources 6.7 (Huang et al. (2009)). Annotations on the human genome were used because information on the pig genome was found to be limited.

SNP by genetic background interaction. SNP effects may differ depending on the genetic background in which the alleles are present. To estimate effects in different genetic backgrounds the comparison of purebred and crossbred population is particularly useful, both because 50 or 75% of the genetic background can be replaced within 1 or 2 generations, and because the difference between purebred and crossbred effects are of practical importance in pig breeding (Dekkers (2007)). However, estimating the effect of a SNP allele from breed A, when it is present in a crossbred AB animal requires the purebred origin of alleles in crossbreds to be determined.

A total of 5,692 animals from five populations, three pure lines (herein referred to as breeds) A, C and D, and two crosses, CD (C×D or D×C) and ACD (A×CD), were genotyped with the Illumina PorcineSNP60 Beadchip (Ramos et al. (2009)). Genotypes of SNPs on SSC2 were analysed that had call-rate >0.95 in each breed or cross. No threshold was used for minor allele frequency (MAF). Samples with call-rate >0.98 were included in the analysis. The final dataset contained 2,695 SNPs and 956, 1,816 and 1,918 animals for breeds A, C, and D respectively. The number of crossbred animals included were 324, and 241 for CD and ACD respectively.

To determine the breed origin of alleles in crossbreds we wanted to phase the haplotypes in the crossbred, and subsequently determine the breed origin of the haplotypes. Pedigree-based phasing methods were not suitable, because the parents of genotyped crossbreds were not included in the genotype data. This corresponds to a common situation in real breeding programs where the pedigree of crossbred animals are not known and several generations may separate the genotyped purebred and crossbred animals. LD based phasing methods were also not suitable because haplotypes within a LD block are often common between several pig breeds (e.g. Hidalgo et al. (2014)).

Long range phasing (Kong et al. (2008)) overcomes both the issues of lacking pedigree and common haplotypes between breeds. Genotype data from the 5 populations was combined in one dataset and analyzed with AlphaPhase software version 1.1 (Hickey et al. (2011)). AlphaPhase was run without pedigree information, and allowing 1% genotype errors and 1% disagreement between genotype and haplotypes. Number of surrogates and percentage of surrogate disagreement were both set to 10. AlphaPhase assigns two haplotypes to each animal which were processed using custom functions in R (R Core Team (2013)). First, the unique haplotypes were identified among the pure breeds. A haplotype was considered to originate in a specific breed if >90% of its copies were observed in that breed, otherwise the haplotype origin was set to ambiguous. Second, the haplotypes of crossbreds were matched to the haplotypes assigned to the different breeds. Alleles (0 and 1) carried on the crossbred haplotypes were then assigned their purebred origin.

Allelic origin assignment was done with a range of settings for core and tail length in AlphaPhase and both offset and non-offset analyses were applied. Core lengths ranged from 150 to 350 SNPs in steps of 50, and tail lengths were 50, 100 or 200 SNP. Allelic origin assignments were summarized from 18 different AlphaPhase analyses, using custom functions in R (R Core Team (2013)). Each allele at each SNP received 18 breed origin assignments, or fewer when some of the AlphaPhase analyses resulted in an ambiguous haplotype origin. The most frequent breed assignment was considered the true origin for each allele.

Results and Discussion

SNP by environment interaction. Changes in the environment are likely to put different requirements on the abilities of the animal and therefore we expect that the role of an individual gene, and the effects of genetic variation associated with SNPs near this gene, will change. The genomic reaction norm model (Silva et al. (2014)) did not estimate the SNP effects directly, but first estimated breeding values using the **G** matrix and subsequently SNP effects were back solved for each level of HYS. The overlap between the top 1% (462) of SNP in different HYS levels was high for similar HYS while 47 SNPs were consistently present in the top 1% for all HYS levels ranging from 10 to 22 TNB. These 47 SNPs had relatively high estimates for their effects in all HYS, and as expected the slope of their reaction norm was small.

We now investigated whether the SNP with the largest slopes were near candidate genes that can be related to G×E or robustness. The absolute SNP effects ranged from 0 to 0.0026 across different HYS, and estimates for

the absolute values of slope ranged from 0 to 0.00023 additional piglets per litter per increase of HYS with 1 piglet per litter. The 100 largest absolute values for slope ranged from 0.00010 to 0.00023 (Figure 1). Most of these SNPs, 86, had effects in opposite directions for different HYS environment. While the maximum SNP effect of 0.0026 was not found among these 100 SNPs, the effects did reach values near 0.002 in some HYS while the effects were near 0 in other HYS for the same SNP. The top 100 SNPs for slope were found distributed over the whole genome, ranging from 17 SNPs on SSC1 to 2 SNPs on both SSC9 and SSCX. Some clustering of SNPs was observed, with 24 SNPs within 2 Mbp, and 48 within 5 Mbp of another top 100 SNP.
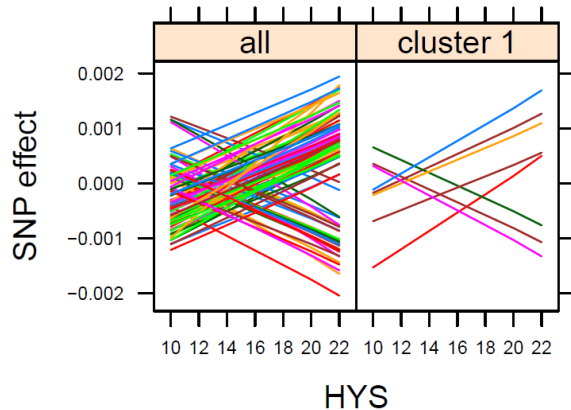

Figure 1. Effect estimates of the 100 largest slope SNP across HYS levels.

Within ±500 kb surrounding the top 100 SNPs, 476 unique gene names were retrieved from Ensembl. Out of the 476 genes, 424 had human annotation records in David, versus only 79 with porcine annotations. The functional clustering of the genes was therefore done with the human annotations. In total 16 clusters were identified, with enrichments scores of 3.54 and 2.72 for groups 1 and 2 that contained 21 (Table 1) and 19 genes respectively. Other groups had enrichment scores of 1.1 or lower. The most enriched functional annotation terms in cluster 1 were UniProt (The UniProt Consortium (2014)) keywords 'defensin' (Fold = 250, P = 1.1×10^{-25}), 'antibiotic' (Fold = 150, P = 7.5×10^{-23}), and 'Antimicrobial' (Fold = 140, P = 1.3×10^{-22}), and GO term (The Gene Ontology Consortium (2000)) 'defense response to bacterium' (Fold = 78, P = 3.2×10^{-20}).

Table 1. Genes and SNPs within functional annotation cluster 1.

Chr	Position	Genes
3	12,847,263	REG3G
5	60,919,433	PLBD1
6	12,847,263	CLEC18A
7	50,260,218	CRISP1, CRISP2, CRISP3, DEFB110, DEFB113, DEFB133
12	358,639	METRNL
12	26,214,886	NXPH3
13	133,415,925	FETUB
17	40,197,933	DEFB115, DEFB116, DEFB119, DEFB121, DEFB123, DEFB124, DEFB125, DEFB128, DEFB129

Chromosome and positions on the Sscrofa10.2 assembly for SNPs in the set with 100 largest slopes. SNPs and nearby genes (±500 Kb) in the first functional annotation cluster.

Responding to bacteria is clearly an interaction with the environment. However, the environmental factor over which the change in SNP effects was estimated was the average production of piglets in the same herd-year-season (HYS), which is not a direct measure of bacterial load or disease. We can speculate that lower average production is correlated with less hygienic conditions, and generally lower levels of husbandry practices. Moreover, a large proportion of the genes in cluster 1 are β-defensins (DEFB) which in mice were shown to be expressed in female reproductive tracts and regulated by the estrous cycle (Hickey et al. (2013)). These gene functions make a functional link to TNB plausible. A higher bacterial load at lower levels of average reproduction could mean that the effects of these genes would become more important at the lower levels of HYS. However, only 2 of the 8 SNPs in cluster 1 have larger absolute effects at lower HYS, 4 have larger absolute effects at high HYS, and 2 have similar absolute effects at both extreme HYS (Figure 1). Having large opposite effects on both ends of the HYS scale may, in part, be the result of using a first order regression model. Fitting a higher order regression may result in a different pattern, allowing smaller values at the extreme HYS.

The majority of the genes in cluster 1 are found close together in two regions on SSC7 and SSC17. When a small number of similar genes are located close together on the genome and all are annotated with a given function then a single hit near these genes will already result in enrichment of this function. Given that the genes in cluster 1 locate in not just one, but two of these regions on separate chromosomes, reduces the chance that we are looking at a spurious result. Even when the enrichment suffers from positional clustering of the DEFB and CRISP genes, the functional clustering was helpful to identify functional candidate genes near the top 100 SNPs. Genes in cluster 1 serve as starting points for investigations towards QTN that interact with environment, potentially affecting robustness.

In the second cluster, with enrichment score 2.72, all 19 genes, except for 2, were homeobox genes, 10 of which were located on SSC18. Homeobox gene products act as transcription factors and can affect the expression of many genes. The functional relationship of these genes with TNB, and how they would interact with the environment or affect robustness is therefore difficult to determine.

SNP by genetic background interaction. To determine the effect of alleles in different genetic backgrounds we need to estimate the effect of the same allele when present in different populations. To be certain that we are measuring the same allele we should estimate the effect of a QTN, but these are largely unknown. If we estimate the effects of SNP markers in different breeds, we

cannot distinguish between changes in LD with the QTN and changes in the QTN effect due to genetic interactions. A solution would be to replace (part of) the genetic background which occurs when crossbred animals are produced. With crossbreeding the effect of a SNP allele can be estimated in a 100% pure breed background and in a background where 50% or 75% is made up by a different breed or breeds. The SNP allele to be estimated will be transmitted to the crossbred together with the same QTL alleles that it is associated with in the pure breed. To enable these estimates we have first worked out a method to identify the alleles from a specific breed when present in crossbred animals.
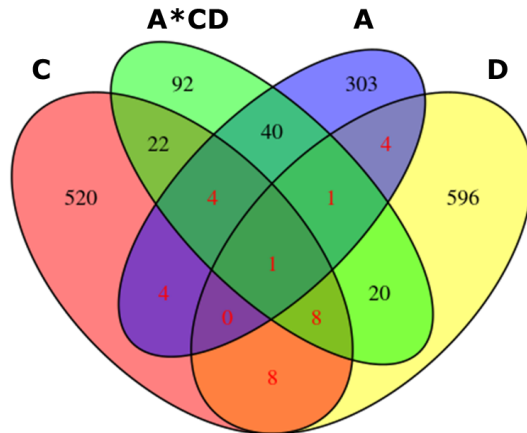

Figure 2. Overlap of unique haplotypes found in the 3 purebreds and the ACD crossbred.

Genotype data from three pure breeds and two crosses were combined in one dataset and analyzed with AlphaPhase1.1 (Hickey et al. (2011)) to produce phased haplotypes for all animals. To be informative for determining purebred origin, a specific haplotype that is found in a crossbred animal should only occur in one of the 3 pure breeds. For instance, from the analysis of a core (genome segment) that was ranging from SNP 401 to 600 on SSC2, 1,623 different haplotypes were observed across the animals of breeds A, C, and D and crossbred ACD (Figure 2). Only a single haplotype was found to be present across each of the three pure breeds, and another 29 haplotypes were shared between two of the breeds. In addition to low haplotype sharing between pure breeds, we need high haplotype sharing between crossbreds and pure breeds. For the same core on SSC2, 270 haplotypes from the 241 ACD animals could be assigned to a single pure breed and 65 were seen in more than one pure breed. Of the remaining 147 haplotypes, 112 were unique to the ACD animals and 35 could not be phased by the software.

To improve the assignment rate from the 56% observed above, a range of settings was applied in 18 different analyses. In addition, the assignment of a haplotype to a pure breed was relaxed by assigning all haplotypes for which at least 90% of the copies were found in a single breed. From the 18 results for each allele in each ACD animal, the most frequent breed assignment for each

allele was used and resulted in between 3% and 100% of an animal's alleles being assigned an origin, with a median of 92%. Out of the 241 ACD animals, 238 had ≥80% of positions assigned and for those animals 47% and 41% of positions had allelic assignments to A+C or A+D respectively. At 10% of positions no assignment was made and 2% of positions had allelic assignments that were inconsistent with the breed composition of ACD animals.

Figure 3 depicts the allelic origin assigned across SSC2 for 20 animals and showed that most ACD animals have a fairly intact chromosome originating from breed A which supplied the paternal chromosome. The other chromosome showed large blocks of green and blue, indicating that origin of alleles is consistently assigned across very long stretches of the chromosome received from the F1 mother. A little over half of the maternal chromosomes show a recombination, with more recombinations towards the end of the chromosomes and fewer near the middle, consistent with the map length and recombination rate being higher in the more distal part of the chromosome (Tortereau et al. (2012)).
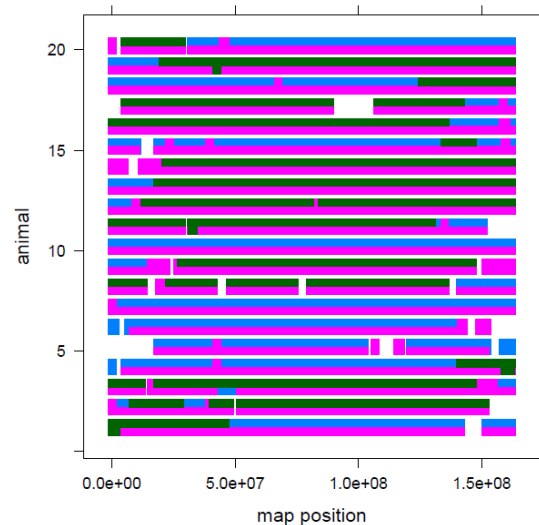

Figure 3. Purebred origin of alleles in 20 ACD crossbred pigs. Two haplotypes per animal are shown with alleles from breeds A (pink), C (green) and D (blue). White regions indicate unassigned allelic origin or regions not covered with SNPs.

The allelic origin results were obtained on real genotype data and accuracy of assignment can therefore not be tested. The approach will need to be applied to simulated data to assess the true accuracy of allelic origin assignment and to test how different factors affect accuracy. However, from the current analyses we conclude that sensible results were obtained for 98% of the genome positions, including 10% unassigned. These 10% unassigned, and the 2% inconsistent assignments leave room for improvement which may be possible by further optimizing the thresholds applied in different steps of the method. To investigate SNP by genetic background interactions, by estimating SNP effects in crossbreds and compare them to their effects in

purebreds, we will need many more genotyped and phenotyped animals than the 241 ACD animals used here.

Cross breeding and SNP effects. Epistatic effect are not important for breeding of pure breeds, as indicated by Crow (2010). Also, while Su et al. (2012) estimated a large epistatic variance they found no improvement in genomic prediction accuracy from including epistatic effects. Of specific interest for pig breeders, however, is the concordance of SNP effects between purebred and crossbred pigs. Simulation results have shown that genomic selection models that take into account the breed origin of alleles in crossbreds can improve the accuracy of EBV but that this only happened in specific circumstances. The differences between breeds should be big enough because the inclusion of breed specific effects in the model comes at a statistical cost (Ibánẽz-Escriche et al. (2009)). In another simulation study, inclusion of dominance effects in the model was shown to outperform the breed specific allele model (Zeng et al. (2013)). Analyses on real data are needed to determine whether the simulation results are good predictors for the value of models that take into account breed specific allele effects, dominance effects, and/or interaction with environment. The differences between breeds may in reality be larger, compared to the simulated data which typically do not simulate interaction effects with genetic background nor the environment.

The interactions of SNPs with environment and with genetic background have been presented here as separate issues. In crossbred animals they are usually confounded. Crossbreds supply a different background for the SNP effects and are typically also kept in different, more variable environmental conditions than the pure breeds. If our concern is to estimate the best SNP effects for prediction of crossbred performance we may not worry about the genetic or environmental origin of the interaction effects. However, if we want to understand the interactions of SNPs with environments, and understand how performance in different genetic backgrounds and different environments comes about, then the disentangling of these effects is needed. The presence of the same cross of animals in many different environments is a benefit for this type of study, but measuring phenotypes may still be a limiting factor.

With genotype data, the performance of crossbreds can be used to estimate SNP effects that can be used to select in pure breeds. With the method presented here the allelic origin of alleles can be determined without the need for tracking pedigree relationships of crossbreds. Moreover, for determining the origin of alleles it is not even necessary to have close relationship between the crossbreds and genotyped purebred animals as long range phasing will work even with distant purebred relatives of the crossbreds.

Conclusions

Evidence of SNP effects that depend on their genetic or environmental context is limited in general, and especially in livestock. It is reasonable to assume that these interactions exist and the few studies that looked into it in real data find considerable interactions. One scenario where SNP by genetic background interactions may be important in livestock, is the difference of SNP effects between purebred and crossbred animals. Estimation of these particular SNP by genetic background interactions requires sufficient genotype and phenotype data and knowledge of the breed origin of alleles. Tracing the purebred origin of alleles in crossbreds was shown to be feasible without close relationships between pure and crossbreds. When crossbred data is collected across a range of environmental values, the genomic reaction norm model allows identification of regions that harbor QTN that interact with the environment. Genotyping of crossbred animals will become more widespread, as selection of purebred animals for crossbred performance appears to have important benefits. Sizeable datasets will therefore be available that allow the estimation of interactions with genetic background and the environment.

Literature Cited

Andersson, L., Haley, C.S., Ellegren, H. et al. (1994). *Science* 263:1771-1774.

Barendse, W. (2011). *BMC Genomics* 12:232.

Calus, M.P.L., Groen, A.F., De Jong, G. (2002). *J. Dairy Sci.* 11:3115-3123.

Chanock, S.J., Manolio, T., Boehnke, M. et al. (2007). *Nature* 447:655-660.

Ciobanu, D.C., Bastiaansen, J.W.M., Malek, M. et al. (2001). *Genetics* 159:1151-1162.

Crow, J.F. (2010) *Phil. Trans. R. Soc. B* 365:1241-1244.

Dekkers, J.C.M. (2007) *J. Anim. Sci.* 85:2104-2114.

de Roos, A.P.W., Hayes, B.J., Goddard, M.E. (2009). *Genetics* 183:1545-1553.

Duchemin, S., Bovenhuis, H., Stoop, W.M. et al. (2013). *J. Dairy Sci.* 96:592-604.

Flicek, M., Amode, M.R., Barrel, D. et al. (2014). *Nucl. Acids Res.* 42:D749-D755.

Groenen, M.A.M., Archibald A.L., Uenishi H., et al. (2012). *Nature* 491:393-398.

Hickey, D.K., Fahey, J.V., Wira, C.R. (2013). *Innate Immun.* 19:121-131.

Hickey, J.M., Kinghorn, B.P., Tier, B., et al. (2011). *Genet. Sel. Evol.* 43:12.

Hidalgo, A.M., Bastiaansen, J.W.M., Harlizius, B., et al. (2014). *BMC Genet.* 15:4.

Huang, D.W., Sherman, B.T., Lempicki, L.A. (2009). *Nat. Prot.* 4:44-57.

Ibánẽz-Escriche, N., Fernando, R.L., Toosi, A., et al. (2009). *Gen. Sel. Evol.* 41:12.

Knap, P.W. (2005). *Aust. J. Exp. Agric.* 45:763-773.

Kong, A., Masson, G., Frigge, M.L. et al. (2008). *Nat. Genet.* 40:1068-1075.

Lillehammer, M., Hayes, B.J., Meuwissen, T.H.E. et al. (2009). *J. Dairy Sci.* 92:4008-4017.

Mackay, T.F.C. (2014). *Nat. Rev. Genet.* 15:22-33.

R Core Team (2013). http://www.R-project.org/.

Ramos A.M., Crooijmans R.P.M.A., Affara N.A. et al. (2009). *PLoS One* 4:1-13.

Sax K. (1923). *Genetics* 8:552–560.

Silva, F.F., Mulder, H.A., Knol, E.F., et al. (2014). *J. Anim. Sci.* in press : doi:10.2527/jas.2013-6486.

Streit, M., Wellmann, R., Reinhardt, F. et al. (2013). *G3* 3:1085-1093.

Su, G., Christensen, O.F., Osterson, T. et al. (2012). *PLoS One* 7:e45293.

The Gene Ontology Consortium (2000). *Nat. Genet.* 25:25-29.

The UniProt Consortium (2014). *Nucl. Acids Res.* 42:D191-D198.

Tortereau, F., Servin, B., Frantz L. et al. (2012). *BMC Genomics* 13:586.

Wilkinson, S., Lu, Z.H., Megens, H-J. et al. (2013) PLoS Genet. 9: e1003453.

Zeng, J., Toosi, A., Fernando, R.L. (2013). *Genet. Sel. Evol.* 45:11.