A rapid method for the identification of epistatic 'dormant' SNPs

A. Reverter[1], J. Henshall[2], L.R. Porto-Neto[1], F. Raidan[1], Y. Li[1], M. Naval-Sánchez[1], S. Dominik[3], S.A. Lehnert[1], K. Meyer[4], Z. Vitezica[5] & A. Legarra[5]

[1]CSIRO Agriculture & Food, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia
Tony.Reverter-Gomez@csiro.au (Corresponding Author)
[2]Cobb-Vantress Inc., Siloam Springs, Arkansas 72761-1030, USA
[3]CSIRO Agriculture & Food, Armidale, NSW 2350, Australia
[4]AGBU/UNE, Armidale, NSW 2351, Australia
[5]INRA/INPT, UMR 1388 GenPhySE, F-31326 Castanet-Tolosan, France

Summary

We present a unique computational approach for the identification of epistatic SNPs based on SNPs with significant yet opposed effects depending on the genetic background. We introduce the mechanical heuristics of the approach based on first, binning the population according to their genomic-estimated breeding value (GEBV) and second, performing genome-wide association studies (GWAS) within each bin. SNPs are deemed to be epistatic if significant but with different signed effects in the GWAS from the most extreme bins containing individuals with the lowest and highest GEBV. We then show that these heuristics are equivalent to a regression of residuals on GEBV. Next, we illustrate our approach with a dataset of 2,111 cattle genotyped for 651,253 SNPs and using yearling weight as the phenotype. We identify 243 epistatic SNPs, and argue that these SNPs are 'dormant' with an additive effect waiting to be 'released' if selection moves the population to either tail of the genetic value distribution.

Keywords: epistasis, genomic selection, genome wide association

Introduction

The availability of high-density SNP genotypes in livestock species allows for the exploration of non-additive effects to a degree not often captured by pedigree relationships alone. In particular, epistasis —the interaction between loci— is thought to play a key role defining the genetic architecture of complex traits (Mackay, 2014). However, exploring all possible SNP to SNP combinations is computationally prohibitively and statistically underpowered. Hence, alternative compromises are being proposed such as the identification of higher-order interactions such as one SNP against the polygenic background (Crawford *et al.*, 2017).

Inspired by these models, here we present a unique computational approach for the rapid identification of epistatic SNPs based on those with significant effect to the phenotype, however with an opposed effect depending on the genetic background of the sampled population.

Materials and methods

GWAS for epistasis: one locus against polygenic background

A typical model is where is estimated as vector with additive polygenic effects. Let and assume that there is an epistatic deviation QTL at position with statistical (not functional) effect and that the epistasis is against the polygenic background. A model for total genotypic value is: (Jannink, 2007), where is a centered vector with for genotypes .

Equivalently, **,** where is a matrix whose diagonal contains the coding of the different genotypes at locus *i*. Thus, can be seen as the regression of the remaining genetic value once the polygenic additive effect has been removed from .

Imagine for instance the epistatic effect is and . For an individual with and carrier of genotype, the epistatic effect is negative: , $2 - 2p = 0.8$, and the total genetic value is . Similarly, for an individual with , the epistatic QTL has no effect; for an individual with , the epistatic effect is positive.

Mechanical heuristics

In lay terms, our proposed approach proceeds in five main steps as follows:

(1) Rank individuals from lowest to highest genomic estimated breeding value (GEBV).
(2) Divide the ranked list in five equally-sized bins with BIN1 containing the 20% of individuals with the lowest GEBVs, BIN2 the next 20% of individuals based on GEBVs, and so on until BIN5 containing the 20% individuals with the highest GEBVs.
(3) Perform a GWAS of SNPs on phenotypes, within bin and with the whole population.
(4) Collect SNPs with significant yet opposed effect in BIN1 and BIN5 and a monotonic pattern of effect from BIN1 to BIN5 (eg. Strong positive, mild positive, zero, mild negative, and strong negative).
(5) Confirm the SNPs collected are not significant in the GWAS with the whole population.

The interpretation of the heuristic is that we try to find the epistatic SNPs that would be significant in extreme populations but are not significant in the current population.

Fast approximate numerical method

The quantity of interest is the regression of on , which can be approximated as follows:

(1) Run a GBLUP with additive effect.
(2) Extract residuals and GEBVs from the output.
(3) For each SNP marker :
    a. Multiply by centered gene contents to obtain
    b. Run a single marker regression to estimate
    c. Obtain a t-test and associated *P*-value from the output.

This approximate method is very fast, but ignores the uncertainty in the estimation of and . It may be used for a fast screening followed by a REML analysis (Jannink, 2007; Crawford *et al.*, 2017) for a subset.

Animals, phenotypes and genotypes

We use a beef cattle dataset of 2,111 Brahman individuals genotyped for 651,253 autosomal SNPs and with yearling weight as the quantitative phenotype (Reverter *et al.,* 2017). In brief,

the dataset included 1,116 bulls and 995 heifers with mean ((±SE) yearling weight of 243.71 (±0.87) kg and 209.73 (±0.97) kg, respectively.

The 651,253 SNP were selected from the ~777K SNP in the BovineHD BeadChip array (Illumina, Inc.) located in autosomal chromosomes and having a minor allele frequency > 1%.

Results and discussions

Figure 1 shows the steps of the mechanical heuristics proposed to identify SNPs with epistatic effect. Using a nominal P-value < 5% from the GWAS within the extreme bins (BIN1 and BIN5) with opposite effect sign, plus a monotonic pattern of effect across bins as well as a P-value > 10% in GWAS using the whole population, the mechanical heuristics identified 243 epistatic SNPs distributed genome-wide. These included 113 and 130 with a "negative to positive" and a "positive to negative" pattern across bins, respectively.

As examples, four of these SNPs, including two of each pattern, are listed in Table 1. Table 1 also lists the effect of a SNP in the coding regions of *PLAG1*, a well-known loci affecting growth and fertility in cattle (Fortes *et al*., 2013). The SNP of *PLAG1* was found to be significant only in the GWAS of the middle bin (BIN3) and in the GWAS of the whole population.

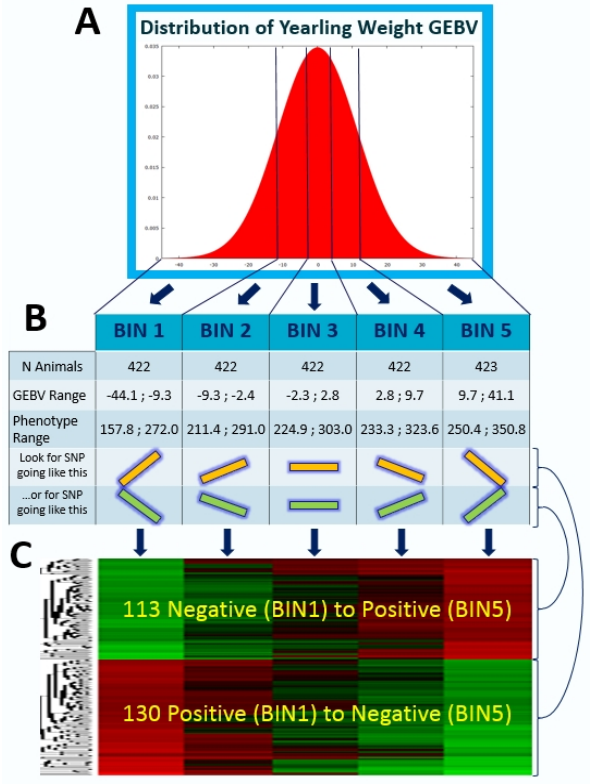

Figure 1. Mechanical heuristic to identify epistatic SNP: (A) Distribution of yearling weight GEBV for 2,111 animals with five equally sized bins clearly demarked, BIN1 to BIN5; (B) Across bins the range of GEBV are by construct non-overlapping, but the range of phenotypes overlap across bins. A GWAS of SNP genotype on phenotype is performed with the intention to capture SNP with significant yet opposed effect in BIN1 and BIN5 and a monotonic pattern of effects across bins; (C) A total of 243 epistatic SNP are found including 113 showing a "negative (BIN1) to positive (BIN5)" pattern, and 130 showing a "positive (BIN1) to negative (BIN5)" pattern

Among the genes listed in Table 1 we highlight *LRIG3* (Leucine-rich repeats and immunoglobulin-line domains protein 3), a body size-related gene found to be under positive selection in a study of five bovine breeds including Brahman (Xu *et al*., 2015). This finding is of most relevance because genes found under selection in a breed comparison study are bound to have little variation in their coding region and/or no additive effect in any given breed, and are only identified as relevant, such as harbouring signatures of selection, in a multi-breed comparison.

Figure 2 illustrates the relationship between the mechanical heuristics and the numerical approximation based on the regression of residuals on GEBV. A correlation of 0.690 was observed between the two approaches. We hypothesize the epistatic SNPs found here as being 'dormant' with an additive effect waiting to be 'released' when selection moves

the population to either tail of the genetic value distribution. Consistent with the argument of Carlborg *et al.* (2006), we further argue that these SNPs provide an answer to the long-standing paradox by which genetic variation does not diminish with selection as fast as theory would anticipate, and instead epistasis is responsible for the release of genetic variation during long-term selection.

Table 1. Estimated SNP effects in the GWAS within BINs and in the whole population: Two examples each of "Negative to Positive" and "Positive to Negative" pattern as well as for a SNP in the PLAG1 coding region. Asterisks indicate significance at P < 0.001.

SNP chr:bp (*Gene*)	BIN1	BIN2	BIN3	BIN4	BIN5	Whole
18:56.5 (*CPT1C*)	-7.58*	-1.38	-0.86	2.67	4.84*	0.60
28:23.3 (*CTNNA3*)	-8.00*	-3.05	-1.93	1.55	5.06*	-0.05
27:1.1 (*CSMD1*)	4.53*	1.18				
5:54.9 (*LRIG3*)	4.86*	-0.30				
14:25.0 (*PLAG1*)	0.74	2.07				

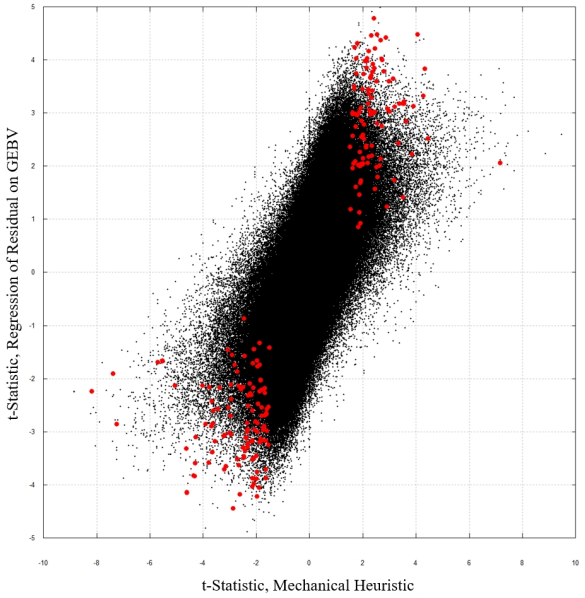

Figure 2. Relationship between the mechanical heuristic and the regression of residuals on GEBV. Highlighted red are the 243 epistatic SNP according to the mechanical heuristics

List of References

Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel & L. Andersson, 2006. Epistasis and the release of genetic variation during long-term selection. Nat Genet. 38(4): 418–420.

Crawford, L., P. Zeng, S. Mukherjee & X. Zhou, 2017. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 13(7): e1006869.

Fortes, M.R., K. Kemper, S. Sasazaki, A. Reverter, J.E. Pryce, W. Barendse, R. Bunch, R. McCulloch, B. Harrison, S. Bolormaa, Y.D. Zhang, R.J. Hawken, M.E. Goddard & S.A. Lehnert, 2013. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. Anim Genet. 44(6): 636-467.

Jannink J.-L., 2007. Identifying Quantitative Trait Locus by Genetic Background Interactions in Association Studies. Genetics 176: 553–561.

Mackay, T.F.C., 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 15(1):22–33.

Reverter, A., L.R. Porto-Neto, M.R.S. Fortes, P. Kasarapu, M.A.R. de Cara, H.M. Burrow & S.A. Lehnert, 2017. Genomic inbreeding depression for climate adaptation of tropical cattle. J Anim Sco. (in press).

Xu, L., D.M. Bickhart, J.B. Cole, S.G. Schroeder, J. Song, C.P. Tassell, T.S. Sonstegard &

G.E. Liu, 2015. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. Mol Biol Evol. 32: 711-725.