Epistasis against background: a source for future genetic variation and a refuge for selection sweeps

Antonio Reverter[1], Laercio R. Porto-Neto[1], Marina Naval-Sánchez[1], John Henshall[2], Fernanda S. S. Raidan[1], Yutao Li[1], Karin Meyer[3], Nicholas J. Hudson[4], Zulma G. Vitezica[5] and Andrés Legarra[5]

[1]*CSIRO Agriculture & Food, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia.*
[2]*Cobb-Vantress Inc., Siloam Springs, Arkansas 72761-1030, USA.*
[3]*Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia.*
[4]*School of Agriculture and Food Sciences, The University of Queensland, Gatton, QLD 4343, Australia.*
[5]*INRA/INPT, UMR 1388 GenPhySE, F-31326 Castanet-Tolosan, France.*

Abstract

We present a unique computational approach for the identification of epistatic single-nucleotide polymorphisms (SNPs) showing significant yet opposed effects depending on genetic background. We introduce the mechanical heuristics of the approach by 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 they have a significant effect but with opposed sign in the GWAS from the most extreme bins containing individuals with the lowest and highest GEBV. We then show that these heuristics can be formally postulated by means of random regression on gene content. We finally propose a very fast approximate method based on a regression of residuals on GEBV. We apply our approach to a dataset of 4,661 cattle from two populations with high-density genotypes and using yearling weight as the test phenotype. We identify epistatic SNPs present in or near genes reported in the context of signatures of selection in multi-breed cattle population studies. These include loci known to be associated with coat color, fertility and adaptation to tropical environments. We argue that these SNPs are 'dormant' with an additive effect waiting to be 'released' – i.e., converted into additive genetic variation – when 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 $\boldsymbol{y} = \cdots + \boldsymbol{W}\boldsymbol{u} + \cdots$ where $\boldsymbol{u}$ is estimated as vector with additive polygenic effects. Let $Var(\boldsymbol{u}) = \boldsymbol{G}\sigma_u^2$ and assume that there is an epistatic deviation QTL at position i with statistical (not functional) effect $(\alpha\alpha)_i$ and that the epistasis is against the polygenic background. A model for total genotypic value is: $\boldsymbol{g} = \boldsymbol{u} + \boldsymbol{i} = \boldsymbol{u} + (\alpha\alpha)_i\boldsymbol{z}_i \odot \boldsymbol{u}$ (Jannink, 2007), where $\boldsymbol{z}_i$ is a centered vector with $\{2-2p, 1-2p, -2p\}$ for genotypes $\{aa, Aa, AA\}$.

Equivalently, $\boldsymbol{g} = \boldsymbol{u} + (\alpha\alpha)_i\boldsymbol{Z}_i^*\boldsymbol{u}$ **,** where $\boldsymbol{Z}_i^*$ is a matrix whose diagonal contains the coding of the different genotypes at locus *i*. Thus, $(\alpha\alpha)_i$ can be seen as the regression of the remaining genetic value once the polygenic additive effect u has been removed from g.

Imagine for instance the epistatic effect is $(\alpha\alpha)_i = -0.2$ and $p = freq(a) = 0.6$. For an individual with $u = 20$ and carrier of aa genotype, the epistatic effect is negative: $(\alpha\alpha)_i z_i u = -0.2 \times 0.8 \times 20 = -3.2$, $2-2p = 0.8$, and the total genetic value is $g = 20 - 3.2 = 16.8$. Similarly, for an individual with $u = 0$, the epistatic QTL has no effect; for an individual with $u = -20$, 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.

Exact method for detection and effect estimation

The exact method is based on a random regression model with two traits (genetic effects): a general additive trait $\boldsymbol{u}$ with variance $\boldsymbol{G}\sigma_u^2$, and a transformation of the epistasis into another additive trait that describes the interaction with the realized genotype of the marker: $\boldsymbol{u}^i = \boldsymbol{u}(\alpha\alpha)_i$, $Var(\boldsymbol{u}^i) = \boldsymbol{G}\sigma_u^2(\alpha\alpha)_i^2$.

The method proceeds by likelihood ratio test of the two alternative hypothesis, H_0: $\boldsymbol{G}_0[1,2] = \boldsymbol{G}_0[2,2] = 0$ versus H_1: $\boldsymbol{G}_0[1,2] \neq 0, \boldsymbol{G}_0[2,2] \neq 0$. For each *i*-th marker the corresponding covariance matrix $\boldsymbol{G}_0$ must be estimated, i.e. by REML using random regression with the Model H_1:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{z}_i\alpha_i + \boldsymbol{W}\boldsymbol{u} + \boldsymbol{W}\boldsymbol{Z}_i^*\boldsymbol{u}^i + \boldsymbol{e}$$

and with the Model H_0:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{z}_i\alpha_i + \boldsymbol{Wu} + \boldsymbol{e}$$

where $\boldsymbol{W}$ is an incidence matrix relating animal to records, $Var\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{u}^i\end{pmatrix}$ is as non-full rank join covariance matrix, $Var(\boldsymbol{e}) = \boldsymbol{R}$ as usual and $\boldsymbol{Z}_i^*$ contains centred gene scores. The regression on gene content $\boldsymbol{z}_i\alpha_i$ corrects for eventual additive (not epistatic) effects of the markers.

After fitting the two models, the likelihood ratio test of the competing models is distributed as a mixture of 0 and 1 degrees of freedom chi-square, from which *P* values can be obtained. In addition, from the estimated covariance matrix $\boldsymbol{G}_0$ estimated under H_1, the estimated epistatic effect can be obtained as:

$$\widehat{(\alpha\alpha)}_i = \widehat{\boldsymbol{G}}_0[2,2]/\widehat{\boldsymbol{G}}_0[1,2].$$

Fast approximate numerical method

The quantity of interest is the regression of $\boldsymbol{y}$ on $\boldsymbol{Z}_i^*\boldsymbol{u}$, which can be approximated as follows:

(1) Run a GBLUP with additive effect.
(2) Extract residuals $\hat{\boldsymbol{e}}$ and GEBVs $\hat{\boldsymbol{u}}$ from the output.
(3) For each SNP marker i:
   a. Multiply $\hat{\boldsymbol{u}}$ by centered gene contents to obtain $\boldsymbol{Z}_i^*\hat{\boldsymbol{u}}$
   b. Run a single marker regression $\hat{\boldsymbol{e}} = \mathbf{1}\mu + (\alpha\alpha)_i\boldsymbol{Z}_i^*\hat{\boldsymbol{u}} + \boldsymbol{\epsilon}$ to estimate $(\alpha\alpha)_i$
   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 $\hat{\boldsymbol{e}}$ and $\hat{\boldsymbol{u}}$. 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 previously reported (Reverter *et al.*, 2017) beef cattle dataset consisting of 2,111 Brahman (BB) and 2,550 Tropical Composite (TC) individuals genotyped for 651,253 and 689,818 autosomal SNPs, respectively. We use yearling weight as the quantitative phenotype of interest. The average (±SE) YWT for BB and TC was 227.70 (±0.75) kg and 247.07 (±0.87) kg, respectively.

Results and discussions

Figure 1 shows the steps of the mechanical heuristics proposed to identify SNPs with epistatic effect. Using a nominal $P < 0.05$ 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 $P > 0.10$ in the GWAS using the whole data, we found 243 and 143 epistatic SNPs in the BB and TC population, respectively.

For each population, six of these SNPs, including three of each pattern (positive to negative, and negative to positive), 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) for the BB population and in the GWAS of the whole data in both populations.

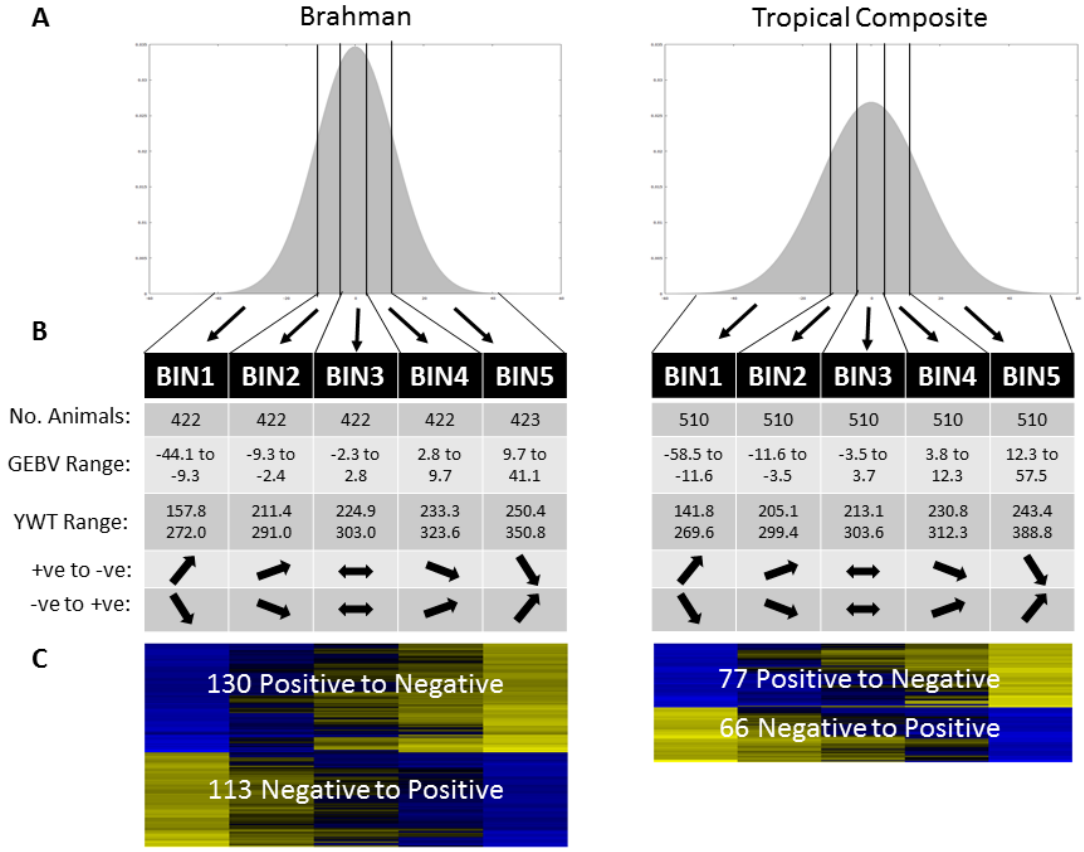

Figure 1. Schematic flowchart of the mechanical heuristic to identify epistatic SNP: (A) Distribution of yearling weight GEBV for 2,111 BB and 2,550 TC cattle with five equally sized bins clearly demarked, BIN1 to BIN5; (B) Across bins the range of GEBV are by construct non-overlapping, but the range of phenotypes overlap across bins. A GWAS of SNP genotype on phenotype is performed with the intention to capture SNPs with significant yet opposed additive effects in BIN1 and BIN5, and with a monotonic pattern of effects across bins; (C) We find 243 and 143 epistatic SNPs in the BB and TC population, respectively.

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). Similarly, the gene *GRIK1* (*Glutamate Receptor, Ionotropic, Kainate 1*) at 5.39 Mb of BTA1 is involved in differentiation of osteoclasts and osteoblasts activities in bone as has been reported as being under selection in tropical adaptation in a comparison of European, African and Zebu cattle (Flori et al., 2012). Finally, located on 28.5 Mb of BTA16, *CNIH3* (*Cornichon homolog 3*) has been found to contain the most significant signal of recent positive selection in dairy and beef cattle (Utsunomiya et al., 2013).

These findings are 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.

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. Mechanical Heuristics: Estimated SNP effects in t-statistic units (estimated effect divided by standard error) in the GWAS within BINs and across the whole population: Three examples each of "Negative to Positive" and "Positive to Negative" pattern as well as for a SNP in the PLAG1 coding region for Brahman and Tropical Composite populations. Asterisks indicate significance at $P < 0.001$.

SNP Chr:Mb (Gene)	BIN1	BIN2	BIN3	BIN4	BIN5	Whole
			Brahman			
18:56.5 (CPT1C)	-7.58*	-1.38	-0.86	2.67	4.84*	0.60
23:50.0 (PRPF4B)	-4.50*	0.91	1.39	2.14	3.58*	1.28
28:23.3 (CTNNA3)	-8.00*	-3.05	-1.93	1.55	5.06*	-0.05
4:71.4 (OSBPL3)	5.95*	1.45	-0.50	-1.89	-6.01*	0.29
5:54.9 (LRIG3)	4.86*	-0.30	-0.79	-2.05	-3.94*	1.18
27:1.1 (CSMD1)	4.53*	1.18	-0.66	-0.69	-2.37*	0.58
14:25.0 (PLAG1)	0.74	2.07	3.46*	2.15	2.02	4.71*
			Tropical Composite			
14:84.3 (SNTB1)	-4.21*	0.25	1.69	1.93	3.42*	0.64
16:28.5 (CNIH3)	-3.39*	-2.00	0.65	1.10	2.74*	0.52
23:15.3 (FOXP4)	-6.20*	-1.94	-0.76	1.87	5.14*	-0.26
1:5.39 (GRIK1)	2.90*	0.74	0.16	-0.83	-2.98*	1.58
10:75.6 (KCNH5)	3.56*	2.07	1.00	-0.12	-3.63*	-0.54
22:57.3 (PPARG)	3.49*	1.12	-0.09	-2.66*	-2.81*	1.01
14:25.0 (PLAG1)	-0.08	2.83*	2.47	1.02	0.44	6.16*

Acknowledgments

This work was undertaken while Antonio Reverter and Andrés Legarra were the recipients of a CSIRO/INRA Linkage Grant during the 2016/17 period. Andrés Legarra and Zulma Vitezica acknowledge financing from INRA SelGen metaprogram (project EpiSel) and project Genopyr (FEDER funds and Region Aquitaine, France).

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.

Flori L, Gonzatti MI, Thevenon S, Chantal I, Pinto J, Berthier D, et al. 2012. A Quasi-Exclusive European Ancestry in the Senepol Tropical Cattle Breed Highlights the Importance of the slick Locus in Tropical Adaptation. PLoS ONE. 2012;7(5):e36133.

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 Sci. 95(9):3809-3821.

Utsunomiya YT, Pérez O'Brien AM, Sonstegard TS, Van Tassell CP, do Carmo AS, Mészáros G, et al. 2013. Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods. PLoS ONE. 2013;8(5):e64280.

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.