Journal of Genetic Syndromes & Gene Therapy
ISSN: 2157-7412

Hartwig, J Genet Syndr Gene Ther 2013, 4:9
DOI: 10.4172/2157-7412.1000189

Review Article Open Access

SNP-SNP Interactions: Focusing on Variable Coding for Complex Models of Epistasis

Fernando Pires Hartwig*

Postgraduate Program in Epidemiology, Department of Social Medicine, Faculty of Medicine, Federal University of Pelotas, Brazil

Abstract

Genetic epidemiology is a promising field to identify patterns of disease susceptibility that can be explored in personalized medicine. However, especially for complex traits, the genetic component is likely to be composed of several loci and/or of interactions between them. The last is addressed in this manuscript, which aims to provide an overview of the advantages and disadvantages of statistically-oriented and biologically-oriented approaches for two-SNP interactions. Eight biologically-oriented models of epistasis are discussed, focusing on their implementation, which is exemplified with real data. Additionally, some key technical points (such as reducing statistical power due to multiple testing and use of conceptual considerations) are discussed, and an exploratory step prior to the analysis is proposed to pre-select the models of epistasis to be actually tested. A function (written in R) is provided (under request) to facilitate the implementation of such models (and can be easily modified to implement others). It is stressed that, regardless of the method choice, the biological meaning of the model being tested is critical for correct interpretation of the results.

Keywords: Genetic epidemiology; SNP-SNP interactions; Epistasis; Variable coding; Biological meaning

Abbreviations: SNP: Single Nucleotide Polymorphism; GWAS: Genome-Wide Association Study; OR: Odds Ratio; 95% CI: 95% Confidence Intervals

Introduction

The ability to predict disease risk and/or identify individuals or groups of individuals susceptible to different health conditions lies at the core of epidemiology. Considering epidemiology history, the recognition that genetics have roles in human diseases other than obvious genetic syndromes is relatively recent. The current major causes of burden of disease (i.e., non-communicable diseases, such as cardiovascular diseases, cancer and diabetes) [1], although known to have strong environmental determinants, are considered multi-factorial consequences of rather complex interactions among environmental, social and genetic factors. Therefore, genetic epidemiology may provide substantial contributions in identifying individuals with higher susceptibility to different conditions, which may have positive implications for health both at the clinical and even at the population level [2,3]. It is important to highlight that genetic epidemiology can also be used for robust causal inference from associations between a given exposure-outcome pair. Such application of genetic epidemiology, so-called Mendelian randomization, relies on the use of appropriate genetic factors that somehow "mimic" the exposure status and has significant strengths (although also particular limitations) in the context of instrumental variables [4].

Advances in the understanding of the human genome, in technologies for DNA analysis (especially regarding assays for studying human genetic variation) and in statistical thinking that incorporates biological aspects of genetics resulted in high-throughput technologies that allow association studies to be performed on a genome-wide scale. Of these, genome-wide association studies (GWAS) are possibly the most prominent example, where a panel of single nucleotide polymorphisms (SNPs) is genotyped in platforms suited for millions of SNPs at relatively affordable costs [5]. However, a major drawback of the majority of studies in genetic epidemiology (especially of studies involving multiple SNPs) is that they are ultimately focused on single-SNP associations with phenotype [6]. Although convenient, such approach is likely to underestimate the roles of genetics in human diseases by disregarding not only the joint effect of multiple loci but the complex interaction network between them. This might be one of the reasons why the expectations of genetic studies in human health were not meet so far [7,8]. Indeed, the approach of genomic prediction has been proposed to address the task of analyzing the roles of multiple SNPs in combination in complex traits [9]. In this manuscript, however, the focus will be on two-SNP interactions.

SNP-SNP interactions in genetic epidemiology studies

An aspect of genetic studies that has been recently receiving more attention in candidate gene approaches is the assessment of SNP-SNP interactions. Since in such studies normally a small number of SNPs is genotyped, testing for interaction is less cumbersome and easier to interpret. In addition, SNP-SNP interaction studies (when plausible and well-conducted) are more likely to meet the expectation of identifying "genomic hotspots" for human diseases than studies that disregard such interactions. Since focusing on few specific genomic regions is much more feasible to have implications for the so-called genomic medicine (although genome-scale methods are becoming increasingly popular and accessible), the correct study of SNP-SNP interactions is of significant relevance. A further positive aspect of SNP-SNP interactions studies is the possibility of statistical modeling by several types of regression techniques, with straightforward implementation of interaction (or effect modification) analysis. Intuitively, the simplest scenario for a SNP-SNP interaction consists of studying two SNPs. Although relying on only two loci, such analysis can be powerful when a true interaction exists. Moreover, the rationale discussed below can

***Corresponding author:** Fernando Pires Hartwig, Postgraduate Program in Epidemiology, Department of Social Medicine, Faculty of Medicine, Federal University of Pelotas, Brazil, Tel: (5553)81347172; E-mail: fernandophartwig@gmail.com

Received August 24, 2013; **Accepted** October 14, 2013; **Published** October 24, 2013

Citation: Hartwig FP (2013) SNP-SNP Interactions: Focusing on Variable Coding for Complex Models of Epistasis. J Genet Syndr Gene Ther 4: 189. doi:10.4172/2157-7412.1000189

be transposed to interactions involving more than two SNPs (although high-order analyses of interaction might require specific techniques to reduce data complexity [10]).

Interaction between two or more SNPs (or genetic loci in general) is commonly referred to as epistasis [11]. Different epistatic effects have been observed in a variety of species, which are useful to define biologically-intuitive models of interaction between two SNPs. This is somehow different than interaction models that would be included in a regular statistical analysis under a regression framework, for instance. Statistically, it would be intuitive and straightforward to carry out an analysis of epistasis as follows: *outcome ~ SNP1 + SNP2 + SNP1:SNP2* (where "~" denotes "described by", "+" denotes the inclusion of independent variables, and ":" denotes "interaction"), using the appropriate regression model. Before doing so, however, it is necessary to define the genetic models of the SNPs themselves before proceeding with the interaction analysis. The coding for each of the five most common genetic models (with genotypes labeled "0" corresponding to the reference category for a given model) for associations involving a single SNP is shown in Table 1. Briefly, the genotypic model assumes that each genotype has independent effects; the overdominant model assesses whether the heterozygous genotype is associated with the outcome when compared with homozygous genotypes. In the dominant model, carrying one or two copies of the variant allele has the same functional implications, while, in the recessive model, such effect would occur only in homozygous individuals for the variant allele. The additive model (sometimes referred to as log-additive for regression frameworks such as binomial or Poisson) assumes that the effect of adding 1 copy of the variant allele is the same across the range of possible values.

A statistically-intuitive approach for SNP-SNP interactions

Given two diallelic SNPs in loci *A* and *B*, the first step of the analysis would be to perform single-SNP associations with the trait. For this, the choice of which genetic models to include can be based on previous studies and conceptual motivations, as well as to avoid studying models that give similar results. Assuming no criteria whatsoever, all five models could be studied for each SNP, resulting in a total of 10 hypothesis tests. While this is a reasonable number, if the same principle is to be applied for SNP-SNP interaction, a total of 25 different models would be tested. In addition to type-I error inflation due to multiple testing (which could be easily solved), the main complication of this approach is that many of these models would be of difficult interpretation (if making any sense at all). Again, prior evidence and conceptual aspects could be considered to define the genetic model of each SNP to perform the interaction analysis. Moreover, results from the single-SNP analysis could be used, such as taking the most significantly associated model(s) of each SNP. Nevertheless, it is often the case that the interaction under study has not been previously investigated, and it is quite possible that none of the single-SNP models resulted in significant or interesting associations (which would be, in fact, a motivation to perform an interaction analysis).

A statistically-intuitive solution would be to assess interaction using models composed of the same coding (i.e., under the same genetic model) for both SNPs. This is, in fact, what some statistical packages (such as "*SNPassoc*" package for R environment for statistical computing [12]) offer as an analysis of epistasis. Although such interaction can be easily assessed by including an interaction term in a regression model, the equivalent variable coding to achieve the same result is shown (for clarity) in Table 2. Evidently, such interaction models are correct in the statistical sense, but this does not guarantee

Genotypes*	Genetic models†				
	Genotypic	Overdominant	Dominant	Recessive	Additive
AA	0	0	0	0	0
Aa	1	1	1	0	1
Aa	2	0	1	1	2

*A: wild type allele at locus*A*; a: variant allele at locus*A*; †Each column represents the coding system for defining five distinct genetic models

Table 1: Variable coding to define five genetic models for a single SNP.

Combined	Models of statistical interaction†				
genotypes*	Gen-Gen	Over-Over	Dom-Dom	Rec-Rec	Add-Add
AA/BB	0	0	0	0	0
AA/Bb	1	0	0	0	0
AA/bb	2	0	0	0	0
Aa/BB	3	0	0	0	0
Aa/Bb	4	1	1	0	1
Aa/bb	5	0	1	0	2
aa/BB	6	0	0	0	0
aa/Bb	7	0	1	0	2
aa/bb	8	0	1	1	4

*A, B: wild type alleles at loci *A* and *B*, respectively; a, b: variant alleles at loci *A* and *B*, respectively; †Each column represents a variable coding system that is equivalent to including an interaction term in a regression analysis having both SNPs in the specified genetic models; Gen-Gen: Genotypic-Genotypic; Over-Over: Overdominant-Overdominant; Dom-Dom: Dominant-Dominant; Rec-Rec: Recessive-Recessive; Add-Add: Additive-Additive; Of these, only the Add-Add model is actually quantitative (the numbers in the other models represent categories).

Table 2: Statistically-intuitive models of SNP-SNP interactions.

that their interpretation is biologically meaningful (this is particularly true for the "Additive-Additive" model, which is not included in the function for epistasis analysis provided by the "*SNPassoc*" package). In more general terms, such interaction models provide analysis of the form "comparing individuals that have (for example) at least one variant allele of each SNP (i.e., Dominant-Dominant model) with the rest". It does not provide complex interaction systems more typical of epistatic models. So, although statistically intuitive and correct, models coded in this way are not likely to suffice to capture interaction in the biological sense.

Defining biologically-oriented complex models of epistasis by adequate variable coding

Differently than statistically-oriented SNP-SNP interaction analysis, obtaining biological-oriented complex models of epistasis cannot be achieved by, for example, simply including an interaction term in a regression model. Rather, the correct variable coding has to be used to establish such models, therefore being a task to the investigator. Although this provides flexibility to test different models, it is important to define biologically-plausible and interpretable models (otherwise a statistically-oriented approach would be preferable since it has a defined criterion). In addition, translating biological knowledge into a coding system might not be straightforward in some instances. In Table 3, the coding system required to define eight distinct complex models of epistasis is provided. The interpretation of each model and the rationale of its coding system are described below.

Dominant epistasis (1): In this model, one SNP has a dominant effect given that the genotype of the other is homozygous for the wild type allele. However, the presence of at least one variant allele in the other SNP overcomes such effect in a dominant fashion. According to column 1.1, the genotypes where there is at least one variant allele of

Combined	Models of epistasis†										
genotypes*	1.1	1.2	2.1	2.2	3.1	3.2	4	5	6	7	8
AA/BB	0	0	0	0	0	0	0	0	0	0	0
AA/Bb	1	2	0	0	0	0	1	0	1	0	1
AA/bb	1	2	1	2	0	0	1	1	1	1	2
Aa/BB	2	1	0	0	0	0	1	0	1	0	1
Aa/Bb	2	2	0	0	0	0	1	0	2	0	2
Aa/bb	2	2	1	2	1	0	1	1	2	1	3
aa/BB	2	1	2	1	0	0	1	1	1	1	2
aa/Bb	2	2	2	1	0	1	1	1	2	1	3
aa/bb	2	2	2	2	1	1	1	1	2	2	4

*A, B: wild type alleles at loci *A* and *B*, respectively; a, b: variant alleles at loci *A* and *B*, respectively; †1: Dominant epistasis; 2: Recessive epistasis; 3: Dominant and recessive epistasis; 4: Double dominant epistasis without cumulative effect; 5: Double recessive epistasis without cumulative effects; 6: Double dominant epistasis with cumulative effect; 7: Double recessive epistasis with cumulative effects; 8: Quantitative; Of these, only the quantitative model is actually quantitative (the numbers in the other models represent categories). Models that have two versions (e.g., 1.1 and 1.2) are the cases where which SNP corresponds to the "A" hypothetical locus and which corresponds to the "B" hypothetical locus is not redundant.

Table 3: Coding system for eight different epistatic models for two-SNP interactions.

SNP B have the same functional significance (coded as "1"), given that the homozygous wild type genotype is observed for SNP A. If genotypes "Aa" or "aa" are observed, then a different functional effect, which overcomes (i.e., is dominant over) the effects of SNP B (regardless of its genotype) is assumed (coded as "2"). As shown in Table 3 (columns 1.1 and 1.2), the choice of which SNP is dominant over the other (if SNP A is dominant over SNP B or SNP B is dominant over SNP A) changes the coding for some categories of combined genotypes.

Recessive epistasis (2): This model is very similar to the previous one, except that the effect under consideration is the recessive. This model assumes that one SNP has a recessive effect given that the genotype observed for the other SNP is not the homozygous variant. In column 2.1, SNP B has a recessive effect if the genotype for SNP A is either "AA" or "Aa" (corresponding to assigning "1" for "AA/bb" and "Aa/bb" combinations of genotypes). However, if the homozygous variant genotype is observed for SNP A (i.e., "aa"), a different functional effect is assumed (coded as "2"), regardless of the genotype observed for SNP B. Similarly to the dominant epistasis model, it is shown in Table 3 (columns 2.1 and 2.2) that the choice of which SNP has a recessive effect that does not depend on the genotype observed for the other changes the coding for some categories of combined genotypes.

Dominant and recessive epistasis (3): According to this model, one SNP has a dominant effect only if the genotype observed for the other is the homozygous variant. In a simpler (although less technical) view, it is similar to the notion that "the recessive effect of one SNP is to allow the other to have a dominant effect". In column 3.1, it is observed that SNP A has a dominant effect given that the genotype "bb" is observed (resulting in the coding of "1" for the combined genotypes "Aa/bb" and "aa/bb"). As shown in Table 3 (columns 3.1 and 3.2), the choice of which SNP has a dominant effect given that the genotype observed for the other is the homozygous variant changes the coding for some categories of combined genotypes.

Double dominant epistasis without cumulative effect (4): This is a very simple epistatic model regarding its coding and interpretation. It assumes that both SNPs have a dominant effect, so the presence of at least one variant allele in at least one of the SNPs is both sufficient and necessary to configure a functional effect, which is the same regardless of the number of variant alleles and in which SNP they occur. This idea can be easily transposed into a coding system by assigning "1" when at least one variant allele is observed.

Double recessive epistasis without cumulative effect (5): This model is also very simple and similar to the previous one. The difference is that it assumes a recessive effect for both SNPs, so observing the homozygous variant genotype in at least one of the SNPs is both sufficient and necessary to produce a functional effect. The respective coding system is simply to code a genotype combination involving at least one homozygous variant genotype as "1".

Double dominant epistasis with cumulative effect (6): As the name suggests, this model is very similar to model 4. The difference is that the presence of at least one variant allele for both SNPs is assumed to have a different effect than if at least one allele is observed for only one of the SNPs (hence the "cumulative effect"). Regarding the coding system, the only difference (when compared to the respective model without cumulative effect) is that a different effect (coded as "2") is assigned to genotype combinations where at least one variant allele in both SNPs is observed ("Aa/Bb", "Aa/bb", "aa/Bb", "aa/bb").

Double recessive epistasis with cumulative effect (7): Again, the name of this epistatic model is suggestive: this model is very similar to model 5. The difference is that the presence of the homozygous variant genotype for both SNPs is assumed to have a different effect than if only one of the SNPs presents the homozygous variant genotype. Regarding the coding system, the only difference (when compared to the respective model without cumulative effect) is that a different effect (coded as "2") is assigned to the double recessive genotype (i.e., "aa/bb").

Quantitative (8): This model of epistasis is also very simple, and the only one (among the models proposed here) which the coding system is actually numeric, rather than representing categories or factors. It assumes that the effect consists of a monotonic function where the independent variable is the number of variant alleles present in the genotypes observed, which can be clearly seen in Table 3 (column 8). Importantly, this model makes no distinction regarding the origin of the variant allele(s) that result in the respective coding. For example, the coding for the genotype combinations "AA/bb", "Aa/Bb" and "aa/BB" is 2, despite the differences regarding the nature of this total of 2 variant alleles.

Using complex models of epistasis: an example with real data

The theoretical considerations regarding SNP-SNP interactions presented are intended not only to assist genetic researches to use biologically-oriented models of epistasis, but also to provide the basic rationale for the correct understanding of what is actually performed by including interaction terms in regression models (which can surely be useful if correctly interpreted) regarding SNP-SNP interactions. In addition, the rationale exposed may also assist researches to develop a coding system to obtain an epistatic model not covered in the present manuscript. As a further contribution, the coding system proposed for the different types of epistasis is illustrated in an example using real data. The coding was performed using a function written in R [13], which will be provided, under request, to anyone who is interested in using it. Importantly, the function writing focused on simplicity to facilitate interpretation and modification to include genetic models not covered here.

The data was obtained from a manuscript published in 2011 about the associations between two SNPs and breast cancer in a case-control study in Taiwan [14]. One of the SNPs studied is a non-synonymous SNP in the *TP53* gene, resulting in an amino acid substitution at codon

72 of the encoded protein (p53) involving Arginine and Proline (hence, the SNP is commonly referred to as *TP53* R72P SNP) [15]. The other SNP is a substitution involving the nucleotides T and G at the position -309 of the *MDM2* gene (*MDM2* T309G SNP), which has been shown to have functional implications for the expression of the encoded protein (MDM2) [16]. It is important to note that *TP53* and *MDM2* are in the same pathway, since MDM2 negatively regulates p53 [17]. Therefore, studying interactions involving R72P and T309G SNPs is a case that fits the "classical" biological sense of epistasis.

A table containing disease status and genotypes for R72P and T309G SNPs can be easily extracted from Table 3 of the manuscript. After doing so, the coding for the eight models of epistasis can be obtained by using the "*epistasis.coding*" R function (available under request). Since the function returns a data frame where columns are the epistatic models and the rows correspond to the original observations, each column of such data frame can be directly used as an independent variable in a regression model. To illustrate this application, a logistic regression model was fitted to the data for each model of epistasis. The results [odds ratio (OR), 95% confidence intervals (95% CI) and p-values of the likelihood ratio chi-squared test (P)] are shown in Table 4.

Final Remarks

Multiple testing × Pre-analysis filtering

A question that immediately arises when several different models for the same general association (in this case, two-SNP interactions and breast cancer) are tested is: should all models be tested or is there a role for subjective pre-selection of models based on conceptual considerations? Unfortunately, there is no universal answer to this question. If the investigation has an exploratory aim, then fitting all the models is reasonable. Indeed, a common situation would be to test all models in a pilot study aiming to define the most likely model(s) of epistasis, and obtain effect size estimates for sample size calculation for the actual investigation to be carried out more adequately. As for conceptual considerations, there is surely room for them. For example: if a pair of SNPs has been observed to interact under a specific model of epistasis in studies involving different outcomes and/or populations, such consistency may well be used to reduce the number of epistatic models to be tested in a future study.

Epistatic	Categories†		
model*	1	2	
	OR (95% CI)		P
1.1	2.90 (1.39-6.46)	2.51 (1.28-5.30)	0.013
1.2	2.24 (1.03-5.16)	2.68 (1.37-5.65)	0.012
2.1	1.40 (0.84-2.35)	1.39 (0.94-2.04)	0.172
2.2	1.46 (0.96-2.23)	1.31 (0.84-2.04)	0.159
3.1	0.99 (0.60-1.64)	-	0.984
3.2	1.25 (0.84-1.87)	-	0.270
4	2.60 (1.34-5.46)	-	0.004
5	1.39 (0.99-1.97)	-	0.061
6	2.62 (1.30-5.64)	2.59 (1.31-5.52)	0.016
7	1.44 (1.00-2.07)	1.15 (0.56-2.31)	0.142
8	1.19 (1.01-1.40)	-	0.041

*The names of the models of epistasis represented by each number were described in Table 3.†These represent the categories (other than '0', which is the reference) formed by the coding system for each model; "-" indicates models that only have one category other than the reference.

Table 4: Results of the crude analysis for the association between R72P and T309G SNPs and breast cancer under eight models of epistasis.

If all or several models of epistasis are tested, then the need to correct for multiple testing (in studies aiming at identifying true associations) arises. Although there are many kinds of corrections, in the majority (if not in all) of them the stringency increases (i.e., the P-value threshold for significance reduces) according to the number of tests performed. Among such methods, the Bonferroni correction, although known to be over-conservative in several situations, is one of the most popular and, given its simplicity [18], will be considered here as an illustration. Assuming a pre-specified α=0.05 (i.e., the significance threshold), the Bonferroni correction consists of defining a new α by diving its original value by the total number of tests performed (which is equivalent to multiplying the P-values by the total number of tests performed). If the P-values shown in Table 4 were Bonferroni-corrected, only the P-value corresponding to model 4 (double dominant epistasis without cumulative effect) would remain statistically significant.

Although in the example above the Bonferroni-correction somehow "facilitated" the choice of the model of epistasis, it is intuitive that correcting for multiple-testing in analyses of epistasis will always be conservative in the sense that only one model is expected to be "true". In this view, the larger the number of models considered in the analysis, more the statistical power to detect the "single true" model is reduced. This rationale further corroborates the importance of pre-selecting the models to be tested prior to the analysis. Although the capacity of doing so largely depends of the availability of literature on a related topic, an exploratory method based on the data itself can be used. Basically, the genotypes of the two SNPs can be combined in a single variable (resulting in nine different categories, as shown in Table 3, unless one or more of the possible pairs of genotypes do not occur in the sample). Then, the reference is defined (normally, the genotype pair composed of the homozygous genotype for both SNPs) and a regression model is fitted having the combined genotypes as the independent variable.

The results can, then, be explored by visual inspection, which facilitates to apprehend the general aspect of the results and get some intuition on what "formal" models of epistasis are more likely to fit the observed pattern. Therefore, based on such visual inspection, at least some epistatic models can be excluded, thus reducing the total number of models of epistasis to be tested (and, consequently, the stringency of the multiple testing correction is also reduced). In Figure 1, the odds ratio (except for the reference category, for which the odds of breast cancer are shown) with 95% confidence intervals are shown for each category of combination of genotypes for R72P and T309G SNPs (in the form R72P/T309G; for R72P, the nucleotides – which are G and C – are shown) as the independent variable and breast cancer as the outcome. Based on the figure, it is evident that the 95% confidence intervals of the ORs for all groups are, in general, above (or poorly including) the nullity, but are not very different among them. Such pattern clearly fits the assumptions of the dominant model of epistasis without cumulative effects (which was, indeed, the model that resulted in the lowest p-value), suggesting that (at least in some situations), the proposed exploratory step prior to the formal analysis may contribute to reducing the inflation of type-I error, thus increasing statistical power by keeping α closer to 0.05 (or any other pre-specified value of acceptable type-1 error). In addition, such exploratory approach can be used to identify patterns that can, in turn, be explored in new models specified by the researcher. Unfortunately, it is unlikely that the exploratory step will always provide such clear indications of the most likely model of epistasis, which directly impacts its capacity to to contribute for maintaining statistical power by excluding unlikely models.

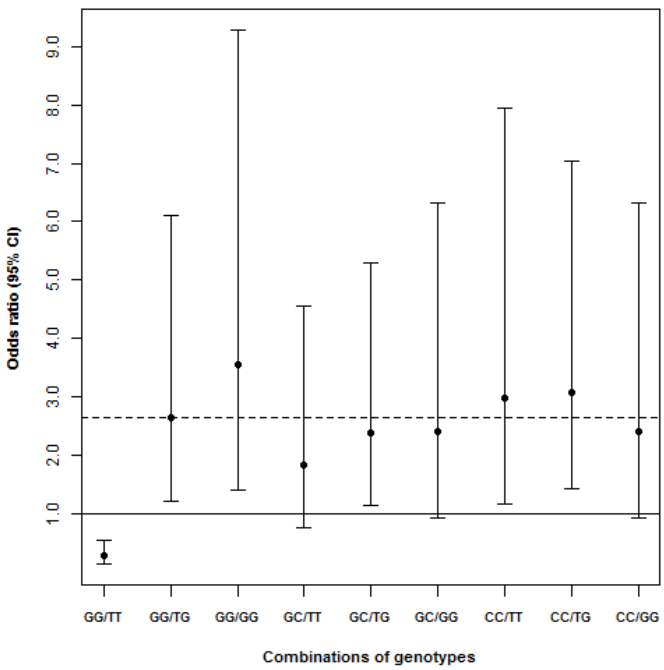

Figure 1: Odds ratio (with 95% confidence intervals) for each combination of genotypes of the *TP53* R72P and *MDM2* T309G SNPs. The genotype "GG/TT" was the reference group, so its plot represents the odds of breast cancer. It is important to make the plot in some "logical" order, since it facilitates its interpretation. In this case, the homozygous wild, heterozygous and homozygous variant genotypes for T309G SNP are shown (from left to right) for each R72P genotype (also following the same order). Based on line that passes through y=1.0, it is easy to notice that the 95% confidence intervals of almost all genotype combinations were above the null (while three poorly incorporated the nullity). The dashed line passes through the mean of the odds ratio values, and facilitates to notice the similarity between the effect estimates.

Statistically-oriented x Biologically-oriented SNP-SNP interactions

Although the focus of this manuscript was to provide a means to investigate SNP-SNP interactions under complex models of epistasis, it should be noted that what was called "statistically-oriented" interaction is not wrong in any sense. What was argued here is that, given its convenience (requiring simply to include an interaction term in a regression analysis), this method could be, in some instances, used without adequate attention to what is actually being analyzed in a biological perspective. For example, it is not essentially wrong to investigate SNP-SNP interactions by including an interaction term between two SNPs, each coded as the dominant model (for example). However, the investigator must be aware of what model of epistasis is actually being analyzed rather than simply interpreting that "interaction", in a general form, is being tested and relying on a significant P-value to conclude that such "general interaction" exists. To facilitate such interpretation, the coding corresponding to including an interaction term between SNPs under different genetic models is shown in Table 2.

In the example above, a biologically-oriented interpretation would be that it is necessary for at least one variant allele to be observed in both SNPs for a functional effect to occur. The warning here is not that this is an implausible model, but that the investigator must be aware that this is the model being investigated. In this regard, if this is the model of choice, then it is surely easier to include an interaction term rather than coding a new variable and adding it to the regression model. Indeed, such practicality is an advantage of statistically-oriented analysis of SNP-SNP interactions. Moreover, when interest lies in analyzing high-order interactions (composed of several SNPs), a statistically-oriented approach might be the only feasible method to perform the analysis. Another advantage is that this approach provides a framework for investigating both main effects of each SNP as well as their interaction. For example: if the interaction consists of including an interaction term between two SNPs (each coded in the dominant model), then the respective coding for the main effects for each SNP is the coding that was used for them in the interaction term (here, the dominant model). In the more complex models of epistasis provided in Table 3, there are no such direct counterparts; in other words, it would be difficult (if possible) to determine how to code the SNPs for studying their main effects in addition to their interaction. The interest of doing so, however, largely depends on the context of the investigation and of the scientific questions to be answered.

Evidently, the methods proposed here for investigating SNP-SNP interactions differ with regards to their application, implementation and interpretation. The fact that these approaches differ does not make them opposites, but rather complements of each other, to be explored adequately by the investigator. Due to the variety of methods presented here for analyzing interactions between two SNPs, the proposed exploratory step may be particularly useful to guide the choosing of models of epistasis to be actually tested. Depending on the directions pointed by the data itself, the investigator can choose between biologically-oriented or statistically-oriented approaches. Regardless of that, being aware of the biological meaning of the model being tested is critical to obtain meaningful results regarding SNP-SNP interactions, as well as for correctly interpreting such results.

Acknowledgments

The author is supported by the Brazilian National Council for Scientific and Technological Development (CNPq)/Brazilian Ministry of Science and Technology (MCT).

References

1. Beaglehole R, Bonita R, Alleyne G, Horton R, Li L, et al. (2011) UN High-Level Meeting on Non-Communicable Diseases: addressing four questions. Lancet 378: 449-455.
2. World Health Organization. *WHO's Human Genetics areas of work*.
3. World Health Organization (2013) *Genes and human disease*.
4. Smith GD, Ebrahim S (2004) Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol 33: 30-42.
5. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363: 166-176.
6. Garrick DJ, Fernando RL (2013) Implementing a QTL detection study (GWAS) using genomic prediction methodology. Methods Mol Biol 1019: 275-298.
7. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7-24.
8. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18-21.
9. Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.
10. Zheng G, Yang Y, Zhu X, Elsto RC (2012) Gene-Gene Interactions, in Analysis of Genetic Association Studies. Springer 235-255.
11. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463-2468.
12. González JR, Armengol L, Solé X, Guinó E, Mercader JM, et al. (2007) SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23: 644-645.

13. Dean CB, Nielsen JD (2007) Generalized linear mixed models: a review and some extensions. Lifetime Data Anal 13: 497-512.

14. Leu JD, Wang CY, Tsai HY, Lin IF, Chen RC, et al. (2011) Involvement of p53 R72P polymorphism in the association of MDM2-SNP309 with breast cancer. Oncol Rep 25: 1755-1763.

15. Whibley C, Pharoah PD, Hollstein M (2009) p53 polymorphisms: cancer implications. Nat Rev Cancer 9: 95-107.

16. Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, et al. (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. Cell 119: 591-602.

17. Jin S, Levine AJ (2001) The p53 functional circuit. J Cell Sci 114: 4139-4140.

18. Laird NM, Lange C (2011) Advanced Topics, in The Fundamentals of Modern Statistical Genetics. Springer 161-174.