frontiers

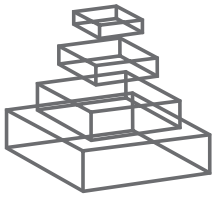
RESEARCH TOPICS

MODELS AND ESTIMATION OF GENETIC EFFECTS

Topic Editors
José M. Álvarez-Castro and Rong-Cai Yang

frontiers in GENETICS

frontiers in ECOLOGY AND EVOLUTION



frontiers

ISSN 1664-8714
ISBN 978-2-88919-444-5
DOI 10.3389/978-2-88919-444-5

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

MODELS AND ESTIMATION OF GENETIC EFFECTS

Topic Editors:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain
Rong-Cai Yang, University of Alberta, Canada

Authors: Ernesto González Torterolo and Ignacio Castro (Náchok)

Ronald Fisher needed to develop elaborate models of genetic effects in order to set the foundations of Quantitative Genetics in his 1918 paper "The correlation between relatives on the supposition of Mendelian inheritance". Since then, many significant implementations have been made to model genetic effects. However, at the verge of one century after Fisher's kick-off, models of genetic effects keep on being discussed and implemented. Indeed, the relatively recent advent of QTL analyses challenged the state of the art of this field by providing researchers the opportunity to obtain and analyze estimates of genetic effects from real data. In this context, the development of this field was not exempt of some polemics, like the debate about the convenience of the functional and the statistical epistasis approaches. This research topic is meant to provide recent developments in models and estimation of genetic effects and to enrich the discussion about how and why models of genetic effects must be further developed and applied.

The articles in this Research Topic shall thus extend, refine and/or provide a refresh look at Fisher's original models of genetic effects and their application to genetic effects estimation and to improve our understanding of evolutionary processes and breeding programs.

Table of Contents

frontiers in
GENETICS

EDITORIAL
published: 12 November 2014
doi: 10.3389/fgene.2014.00396

One century later: dissecting genetic effects for looking over old paradigms

José M. Álvarez-Castro[1]* and Rong-Cai Yang[2]

[1] *Department of Genetics, Universidade de Santiago de Compostela, Lugo, Spain*
[2] *Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada*
**Correspondence: jose.alvarez.castro@usc.es*

Edited and reviewed by:
Samuel A. Cushman, United States Forest Service Rocky Mountain Research Station, USA

Keywords: genetic effects, mathematical modeling, statistical estimation, genetic architecture, environmental effects

The foundation of genetics as a scientific field at the beginning of the twentieth century was not free from controversy. It meant no resolution that the advocates of the Biometric and the Mendelian schools agreed in one thing: the inheritance laws Mendel inferred by studying meristic (discrete) traits did not seem to be compatible with the findings the biometricians had been reporting for continuous (quantitative) variation since the nineteenth century (see Provine, 1971). For providing conclusive evidence against that paradigm, Fisher (1918) developed the foundations of the mathematical models of genetic effects that remain pertinent today, an endeavor in which he developed statistical tools that soon became broadly used beyond genetics.

The genetic effects comprised the core of that theory, but they were initially implemented in those expressions as parameters neither to be estimated nor to actually take any defined numerical values. The most parsimonious hypothesis about genetic effects at that time proposed that the genetic basis of quantitative traits is dominated by the effects of large numbers of genes at which allele substitutions have very small (infinitesimal) and independent (additive) effects on phenotype. This was eventually called the infinitesimal model (see e.g., Bulmer, 1980). Despite the accumulation of evidences suggesting more complex genetic architectures (e.g., Dobzhansky, 1970), the infinitesimal model proved to be a useful paradigm to guide investigation of practical quantitative genetics.

At the time when mapping genetic architectures has moved out the domains of pure fiction (see e.g., Rifkin, 2012), new possibilities for reassessing the adequacy of the infinitesimal model not only reawaken our thirst of knowledge but shall also enable a leap in applicability. It is thus not surprising to witness an increased research effort in updating mathematical and statistical tools for analysing genetic effects, aiming to typify all possible kinds of genetic architectures and their evolutionary implications. We feel grateful for having been able to gather a stimulating account of that update within the current Frontiers Research Topic Issue on Models and Estimation of Genetic Effects.

In the first work in this volume, Gjuvsland et al. (2013) analyse epistasis in genetic networks by focusing on monotonicity as a (correlated) alternative to additivity. Their approach further illustrates that population-referenced (statistical) and non-population-referenced (physiological, functional) genetic parameters are complementary tools in quantitative genetics analyses. The next work, by Le Rouzic (2014), stresses that the evolutionary implications of epistasis are conditioned on whether the interactions follow patterns. He uses the multilinear model to provide practical tools for the detection of such patterns (particularly, directionality) in real data, as well as conceptual keys for aiding the interpretation of the results.

We then move to imprinting, through a work by Álvarez-Castro (2014), who extends the NOIA model to account for that phenomenon and discusses the mathematical properties of the resulting theory in comparison with previous models of imprinting. Further, general procedures for advanced implementation of models of genetic effects are presented in that work. NOIA is also used by Álvarez-Castro and Yang (2012) in the next communication for clarifying the interpretation of the genetic effects defined as average excesses by Ronald Fisher. The interest raised by the publication of that work in Frontiers in Genetics actually triggered the current Research Topic Issue.

A group of papers follows that explicitly account for the environment. Yang (2014) analyses experimental datasets with non-linear functions and addresses some common constraints of the use of linear models to gene by environment interactions. He shows that even under largely linear genotypic responses, strong gene by environment interactions occur because of differences in positions and effects of quantitative trait loci (QTL) between poor and good environments. Marigorta and Gibson (2014) perform simulation studies to tackle the particularities of genome wide association (GWA) human studies. They show that for a wide range of scenarios, cumulative risk of alleles is highly significant despite the lack of evidence for gene by environment interactions, and that increased phenotypic variance after environmental perturbation lowers the statistical power to detect risk alleles in mixed cohorts. The environment of one species may be conditioned by the genome of another, like in the following study by Kodaman et al. (2014) on host-pathogen interactions. They illustrate how pathogens and their human hosts have interacted and coevolved to reduce antagonism and they endorse such information to be incorporated into genetic models to account for the heterogeneity of disease pathology and to avoid dubious conclusions about disease etiology.

The last two communications offer new insights into statistical issues commonly encountered in QTL mapping and GWA studies. Loredo-Osti (2014) provides a bootstrapping procedure to estimate the *p*-values under the mixed-model framework that is applied to QTL mapping when the mapping population consists of recombinant congenic strains, which overcomes a problem concerning the Type I error that had been pointed out in previous

approaches. To conclude our compilation, Dai et al. (2014) address the classic issue of multiple hypothesis tests in the current era of high throughput genomics. They advocate a new (modified Lancaster) procedure that improves the control of the Type I error as compared to the Fisher's combination test as well as to the original Lancaster procedure, whilst maintaining statistical power to detect signals related to biomarkers in pathways.

We also find it worth noting that a couple of interesting works addressing genetic effects have been released during the preparation of this editorial. Wang (2014) provides new developments leading to the same genetic variance decomposition of multiallelic loci under departures from the Hardy-Weinberg proportions that we obtained using NOIA (Álvarez-Castro and Yang, 2011; incidentally, we hereby thank Dr. Wang for pointing out a misprint in one of the values of the applied case we provided in our paper). Varona et al. (2014) also use NOIA for dissecting genetic covariances between individuals in the context of genomic selection. Although this kind of analysis was originally developed under the paradigm of the infinitesimal model, and was specifically designed for accounting for any putative infinitesimal additive genetic signal, it is encouraging that it effectively utilizes innovative models of genetic effects. Finally, we commend the coming publication of a volume devoted to a specific (and important) instance of genetic effects, "Epistasis. Methods and Protocols" Edited by Jason H Moore and Scott M Williams, which can be viewed as a new instalment of the already classical "Epistasis and the Evolutionary Process" (Wolf et al., 2000) and whose author list overlaps with that of this Frontiers Research Topic Issue on Models of Genetic Effects.

We hope the papers in this volume provide a useful compendium of theoretical and statistical developments, data analyses, simulation studies, conceptual contributions and discussion that collectively advance knowledge of genetic architectures and environmental interactions, and their broad implications in evolutionary and population genetics. To better contextualize the consequence of this volume, we recall that the recent Frontiers Specialty Grand Challenge Article of Evolutionary and Population Genetics identifies the integration of genomics, modeling and experimentation as both the most critical challenge and exciting opportunity in advancing our field (Cushman, 2014). We feel that the papers presented in this volume, by showing strong linkages and synergies among modeling, experimentation, genomics and bioinformatics, demonstrate the importance of this kind of integrative research. Updating models of genetic effects is critical to take advantage of the stunning burst of molecular techniques and computing capabilities we are witnessing. Obtaining more general formulations of those models shall enable us to more efficiently characterize genetic architectures and to formulate hypothesis that could better guide experimental and simulation studies. Ultimately, evolutionary and population genetics benefits from the integration of different perspectives, methodologies and scopes of research within it, which in its turn accelerates its integration into a fully-fledged science of evolutionary quantitative genetics.

ACKNOWLEDGMENTS

José M. Álvarez-Castro has been supported by the Autonomous Administration Xunta de Galicia through project EM2014/024 to edit this Research Topic Issue. The authors thank Specialty Chief Editor Samuel A. Cushman for helpful comments.

REFERENCES

Álvarez-Castro, J. M. (2014). Dissecting genetic effects with imprinting. *Front. Ecol. Evol.* 2:51. doi: 10.3389/fevo.2014.00051

Álvarez-Castro, J. M., and Yang, R. C. (2012). Clarifying the relationship between average excesses and average effects of allele substitutions. *Front. Genet.* 3:30. doi: 10.3389/fgene.2012.00030

Álvarez-Castro, J. M., and Yang, R.-C. (2011). Multiallelic models of genetic effects and variance decomposition in non-equilibrium populations. *Genetica* 139, 1119–1134. doi: 10.1007/s10709-011-9614-9

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford: Oxford University Press.

Cushman, S. A. (2014). Grand challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. *Front. Genet.* 5:197. doi: 10.3389/fgene.2014.00197

Dai, H., Leeder, J. S., and Cui, Y. (2014). A modified generalized Fisher method for combining probabilities from dependent tests. *Front. Genet.* 5:32. doi: 10.3389/fgene.2014.00032

Dobzhansky, T. (1970). *Genetics of the Evolutionary Process*. New York, NY: Columbia University Press.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52, 339–433.

Gjuvsland, A. B., Wang, Y., Plahte, E., and Omholt, S. W. (2013). Monotonicity is a key feature of genotype-phenotype maps. *Front. Genet.* 4:216. doi: 10.3389/fgene.2013.00216

Kodaman, N., Sobota, R. S., Mera, R., Schneider, B. G., and Williams, S. M. (2014). Disrupted human-pathogen co-evolution: a model for disease. *Front. Genet.* 5:290. doi: 10.3389/fgene.2014.00290

Le Rouzic, A. (2014). Estimating directional epistasis. *Front. Genet.* 5:198. doi: 10.3389/fgene.2014.00198

Loredo-Osti, J. C. (2014). A cautionary note on ignoring polygenic background when mapping quantitative trait loci via recombinant congenic strains. *Front. Genet.* 5:68. doi: 10.3389/fgene.2014.00068

Marigorta, U. M., and Gibson, G. (2014). A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. *Front. Genet.* 5:225. doi: 10.3389/fgene.2014.00225

Provine, W. B. (1971). *The Origins of Theoretical Population Genetics*. Chicago, IL: University of Chicago Press.

Rifkin, S. (ed.). (2012). *Quantitative Trait Loci (QTL)*. New York, NY: Springer. doi: 10.1007/978-1-61779-785-9

Varona, L., Vitezica, Z. G., Munilla, S., and Legarra, A. (2014). "A general approach for caltulation of genomic relationship matrices for epistatic effects," in *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*. (Vancouver, BC).

Wang, T. (2014). A revised Fisher model on analysis of quantitative trait loci with multiple alleles. *Front. Genet.* 5:328. doi: 10.3389/fgene.2014.00328

Wolf, J. B., Brodie, E. D., and Wade, M. J. (eds.). (2000). *Epistasis and the Evolutionary Process*. New York, NY: Oxford University Press.

Yang, R. C. (2014). Analysis of linear and non-linear genotype × environment interaction. *Front. Genet.* 5:227. doi: 10.3389/fgene.2014.00227

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 October 2014; accepted: 27 October 2014; published online: 12 November 2014.

*Citation: Álvarez-Castro JM and Yang R-C (2014) One century later: dissecting genetic effects for looking over old paradigms. Front. Genet. **5**:396. doi: 10.3389/fgene.2014.00396*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 07 November 2013
doi: 10.3389/fgene.2013.00216

Monotonicity is a key feature of genotype-phenotype maps

Arne B. Gjuvsland[1]*, Yunpeng Wang[2], Erik Plahte[1] and Stig W. Omholt[2,3]

[1] Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås, Norway
[2] Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway
[3] Department of Biology, Centre for Biodiversity Dynamics, NTNU Norwegian University of Science and Technology, Trondheim, Norway

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Arnaud Le Rouzic, Centre National de la Recherche Scientifique, France
Ovidiu Dan Iancu, Oregon Health & Science University, USA

***Correspondence:**
Arne B. Gjuvsland, Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, PO Box 5003, N-1432 Ås, Norway
e-mail: arne.gjuvsland@umb.no

It was recently shown that monotone gene action, i.e., order-preservation between allele content and corresponding genotypic values in the mapping from genotypes to phenotypes, is a prerequisite for achieving a predictable parent-offspring relationship across the whole allele frequency spectrum. Here we test the consequential prediction that the design principles underlying gene regulatory networks are likely to generate highly monotone genotype-phenotype maps. To this end we present two measures of the monotonicity of a genotype-phenotype map, one based on allele substitution effects, and the other based on isotonic regression. We apply these measures to genotype-phenotype maps emerging from simulations of 1881 different 3-gene regulatory networks. We confirm that in general, genotype-phenotype maps are indeed highly monotonic across network types. However, regulatory motifs involving incoherent feedforward or positive feedback, as well as pleiotropy in the mapping between genotypes and gene regulatory parameters, are clearly predisposed for generating non-monotonicity. We present analytical results confirming these deep connections between molecular regulatory architecture and monotonicity properties of the genotype-phenotype map. These connections seem to be beyond reach by the classical distinction between additive and non-additive gene action.

Keywords: genotype-phenotype map, gene regulatory networks, epistasis, variance component analysis, genetic modeling, systems genetics, genetic variance, monotonicity

INTRODUCTION

Quantitative genetics is the major theoretical foundation for genetic studies in production biology, evolutionary biology, and biomedicine. A core concept in quantitative genetics is the genotypic value, the mean observed phenotype for a given genotype. It constitutes the basis for the genotype-to-phenotype (GP) map concept. The shape of a given GP map is typically described by the classical gene action terms: additivity, dominance, and epistasis. Together with genotype frequencies in a given population, the GP map is the basis for decomposing observed phenotypic variance into environmental variance and genetic variance components including additive variance, dominance variance and epistatic variance. This provides the basis for a very successful theory when it comes to predicting selection response and breeding values (Falconer and Mackay, 1996; Lynch and Walsh, 1998) and more recent statistical genetics methods for mapping Quantitative Trait Loci (QTL) (Neale et al., 2008). Quantitative genetics thus provides a mature machinery for predicting the population level consequences of a given GP map, but in order to understand several generic genetic phenomena there is a stated need for new tools for disclosing how the shape of the GP map is determined by underlying biology (Jaeger et al., 2012; Moore, 2012; Gjuvsland et al., 2013).

One such phenomenon is the resemblance between parents and offspring. An explanation in quantitative genetic terms is that the additive variance (V_A) makes up a substantial part of the phenotypic (V_P) and genetic variance (V_G). Hill et al. (2008) showed that in populations with extreme allele frequencies, high V_A/V_G ratios will arise regardless of the shape of the GP map. However, for populations with intermediate allele frequencies a much wider range of V_A/V_G ratios is observed (Wang et al., 2013). In such populations, high V_A/V_G ratios cannot be fully accounted for without considering properties of the GP map. Gjuvsland et al. (2011) showed that a key feature of GP maps that give high ratios of additive to genotypic variance (V_A/V_G), is a monotone (or order-preserving) relation between gene content (the number of alleles of a given type) and phenotype. This led to the hypothesis that the regulatory circuitry of sexually reproducing organisms predominantly predisposes for highly monotone genotype-phenotype maps.

Here we address the above hypothesis by a two-step approach. First we provide methods and software tools for measuring monotonicity of generic GP maps (i.e., sets of genotypic values). Then we use these tools on the data generated by an extensive simulation study of a broad collection of gene regulatory network models. In these network models the steady state expression levels serve as phenotypes and genetic variation is introduced through parameters describing maximal production rates and the shape of the gene regulation function. Such *causally cohesive genotype-phenotype (cGP) models* [see Gjuvsland et al. (2013) and references therein] allow us to identify relationships between regulatory network architecture and properties of the resulting GP maps.

Our results confirm the prediction that the GP maps arising from a wide range of gene regulatory network motifs are in general highly monotone. In addition we show through numerical as well as mathematical analysis that regulatory motifs involving incoherent feed-forward or positive feedback stand out in their capacity to generate non-monotonicity. These relationships between molecular regulatory architecture and properties of the genotype-phenotype map—of substantial relevance to functional genomics in general—are beyond reach by the standard distinction between additive and non-additive gene action.

Our approach can be applied to cGP models of a wide range of biological systems at any level of model complexity. It opens for a systematic study of the monotonicity properties of molecular regulatory structures underlying the whole spectrum of physiological regulation. This suggests that the concept of monotonicity of GP maps can be used to build theory about heredity phrased in terms of molecular mechanism, something which standard genetic concepts and approaches appear to be incapable of.

MODELS AND METHODS

BACKGROUND ON MONOTONICITY OF GP MAPS

To ease understanding we provide a brief recapitulation of the concept of monotonicity (or order-preservation) in GP maps introduced in (Gjuvsland et al., 2011). We consider a diploid genetic model with N biallelic loci (alleles indexed 1 and 2) underlying a quantitative phenotype. A genotype at a single locus k is denoted by $g_k \in \{11, 12, 22\}$. In the case of two loci k and l there are 9 possible genotypes $g_{kl} = g_k g_l \in \{1111, 1112, 1122, 1211, \ldots, 2212, 2222\}$. The general N loci genotype space Γ contains 3^N genotypes $g_1 g_2 \cdots g_N$ (in condensed notation $g_{1:N}$) constructed by concatenating single locus genotypes, $\Gamma = \{g_1 g_2 \cdots g_N \,|\, g_k \in \{11, 12, 22\}, k = 1, 2, \ldots, N\}$.

For any locus k, *the genotypic background*, i.e., the allele composition of all loci *except* k, is $g^{(k)} = g_1 g_2 \ldots g_{k-1} g_{k+1} \ldots g_N = g_{1:k-1} g_{k+1:N}$. For example, if $N = 4$ then $g^{(2)} = 112212$ means that the genotypes of locus 1, 3, and 4 are 11, 22, and 12, respectively. We use the straightforward notation $g_1 g_2 \ldots g_{k-1} 11 g_{k+1} \ldots g_N = g_{1:k-1} 11 g_{k+1:N}$ to indicate a genotype where $g_k = 11$ while the background genotype is arbitrary. We will also use the compressed notation $11 g^{(k)}$ (or generally $g_k g^{(k)}$).

We use the 2-allele content (i.e., the number of 2-alleles) of genotypes to define a partial order on the genotype space Γ (see **Figure 1**, left panel for an illustration). For a particular locus k we order the three genotypes sharing the same background genotype $g_{1:k-1} g_{k+1:N}$ as follows,

$$g_{1:k-1} 11 g_{k+1:N} < g_{1:k-1} 12 g_{k+1:N} < g_{1:k-1} 22 g_{k+1:N} \quad (1)$$

We call this the *partial genotype order relative to locus k*, and it defines a strict partial order on Γ.

A genotype-phenotype map is a mapping G that assigns to each genotype $g \in \Gamma$ a real-valued genotypic value $G(g)$ (the mean trait value for a given genotype). We define monotonicity of G in terms of how it transforms the partial genotype order to the algebraic order of the genotypic values $G(g)$. Without loss of generality we assume that the allele indexes at each locus have been chosen such that $G(1111 \cdots 11)$ is the smallest of all homozygote genotypic values. We call a genotype-phenotype map G *monotone or order-preserving with respect to locus k* if it preserves the partial genotype order relative to locus k, i.e., if,

$$\begin{aligned} G(g_{1:k-1} 11 g_{k+1:N}) &\leq G(g_{1:k-1} 12 g_{k+1:N}) \\ &\leq G(g_{1:k-1} 22 g_{k+1:N}) \end{aligned} \quad (2)$$

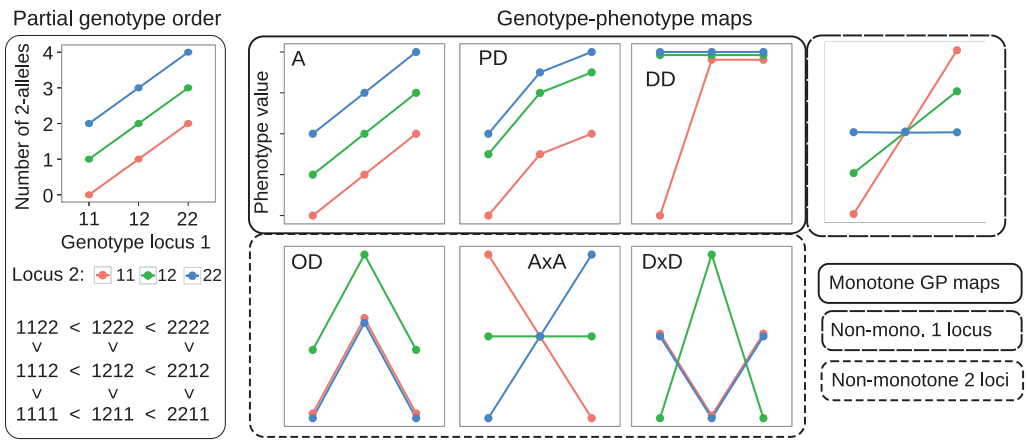


FIGURE 1 | Examples of partial genotype order and genotype-phenotype maps. Left panel: The allele content defines a partial order on genotype space. A two-locus example is shown. The plot at the top displays the genotype at locus 1 (x-axis) and locus 2 (color) vs. the total number of 2-alleles (y-axis) in the two-locus genotype. The resulting partial ordering of genotypes is shown below. **Right panel:** Each lineplot shows the 9 genotypic values (y-axis) for a single GP map, coding of genotype are the same as in the left panel. GP maps that preserve the partial order of genotypes are called monotone. Examples shown are an intra- and interlocus additive map (A), a map showing partial dominance at both loci (PD), and duplicate dominant (DD) epistasis (see Table 1 in Phillips, 1998). GP maps that break the partial order of genotypes are called non-monotone, examples shown are pure overdominance at both loci (OD), additive-by-additive epistasis (A × A) and dominance-by-dominance epistasis (D × D). The rightmost plot shows a GP map that is monotone w.r.t. locus 1, but non-monotone w.r.t. locus 2.

for all genetic backgrounds of locus k. By allowing non-strict inequalities we include GP maps showing complete dominance and complete magnitude epistasis (Weinreich et al., 2005) in the class of order-preserving GP maps. Conversely we call a GP map *non-monotone* or *order-breaking with respect to locus k* if it does not preserve the partial genotype order relative to locus k for all backgrounds. **Figure 1** (right panel) shows classical dominance and epistasis patterns, categorized into monotone and non-monotone GP maps.

STATISTICAL DECOMPOSITION OF GENOTYPE-PHENOTYPE MAPS

Given a genotype-phenotype map G as described above and a corresponding vector of genotype frequencies f in a population, quantitative genetic provides methods for orthogonal decomposition of genotypic values and resulting genetic variance in the population into additive and non-additive (dominance and epistasis) components (Lynch and Walsh, 1998). We performed such statistical decomposition with the function `linearGPmapanalysis` in the R package noia (http://cran.r-project.org/package=noia; Le Rouzic and Alvarez-Castro, 2008) version 0.94.1. We assumed an idealized population where all genotype frequencies are equal ($1/3^N$). In such a hypothetical population the NOIA (Alvarez-Castro and Carlborg, 2007) statistical and functional formulations and the unweighted regression model proposed by Cheverud and Routman (1995) are equivalent. Furthermore, the decomposition of genotypic values is equivalent to decomposing G into a sum of additive and non-additive GP maps, and the genetic variance in this case is simply the variance of the 3^N genotypic values in G. We used the NOIA statistical formulation to decompose a GP map G into its additive and non-additive components, and computed the ratio of additive to total genetic variance V_A/V_G as a measure of how well the additive component describes the original GP map. In case of the illustrative GP maps depicted in **Figure 1**, this gives $V_A/V_G = 1$ for the fully additive GP map A, and $V_A/V_G = 0$ for the pure overdominance (OD) and the pure epistasis (Cheverud and Routman, 1996) maps A × A and D × D.

GENE REGULATORY NETWORK MODELS

Gene expression in eukaryotes is controlled through gene regulatory networks involving numerous regulatory mechanisms [see e.g., Latchman (2005), for details]. Modeling of such gene regulatory networks is well-established, and available modeling frameworks range from coarse-grained descriptions of the topology of genome-wide networks to very detailed mechanistic models describing the dynamics of small networks (De Jong, 2002; Schlitt and Brazma, 2007; Karlebach and Shamir, 2008). In line with a large number of authors we used ordinary differential equations (ODEs) to study a family of generic gene regulatory network models containing three diploid genes X_1, X_2, and X_3, organized as a regulatory system where the rate of expression of each gene can be regulated by the expression level of one or both of the other genes. The wiring of the system is described by a 3×3 connectivity matrix A with elements $A_{kl} \in \{-1, 0, 1\}$. The signs of the elements of A describe the mode of regulation, $A_{kl} = 0$ indicates that X_l is not a regulator of X_k, if $A_{kl} = 1$ then X_l is an activator of X_k, and if $A_{kl} = -1$ then X_l is a repressor of X_k. Gene regulatory systems are often laid out visually as signed directed graphs. There is a one-to-one correspondence between a connectivity matrix and a signed directed graph, two examples are illustrated in **Figure 4**. We used the sigmoid formalism (Mestl et al., 1995; Plahte et al., 1998) in the diploid form (Omholt et al., 2000) where the expression the two alleles of gene k is described by the following ODEs,

$$\begin{aligned} \dot{x}_{k1} &= \alpha_{k1} R_{k1}(y_1, y_2, y_3) - \gamma_{k1} x_{k1}, \\ \dot{x}_{k2} &= \alpha_{k2} R_{k2}(y_1, y_2, y_3) - \gamma_{k2} x_{k2}, \\ y_k &= x_{k1} + x_{k2}, \quad k = 1, 2, 3. \end{aligned} \tag{3}$$

Here α_{ki} is the maximal production rate for allele i of gene X_k, γ_{ki} is the decay rate, while R_{ki} is the gene regulation function (dose-response function). If X_k has no regulators, we assume production is always switched on i.e., $R_{ki} = 1$. If X_k has a single regulator X_l, the gene regulation function is given as $R_{ki}(y_l) = S(y_l, \theta_{lki}, p_{lki})$, where $S(y, \theta, p) = y^p/(y^p + \theta^p)$ if X_l is an activator and $S(y, \theta, p) = 1 - y^p/(y^p + \theta^p)$ if it is a repressor. In both cases the parameter θ_{lki} gives the amount of regulator needed to get 50% of maximal production rate, and p_{lki} determines the steepness of the response. In the case of two regulators X_l and X_j we set $R_{ki}(y_l, y_j) = S(y_l, \theta_{lki}, p_{lki})S(y_j, \theta_{jki}, p_{jki})$, corresponding to the Boolean AND function. Modeling transcription regulation by means of Hill functions and Boolean composition has a long tradition in modeling of gene regulation and is widely used.

With three genes and up to two regulators per gene the number of possible connectivity matrices is 6859. We further required that the system is connected, and that X_3 is downstream to both X_1 and X_2 so either X_1 and X_2 both regulate X_3 directly ($A_{31}A_{32} \neq 0$), or one of them regulates X_3 directly and the other one indirectly ($A_{31}A_{12} \neq 0$ or $A_{32}A_{21} \neq 0$). This reduces the number of distinct connectivity matrices to 3724. Finally, we identified pairs of matrices that are symmetric with respect to interchanging X_1 and X_2 and picked just one matrix from each pair. The resulting 1881 connectivity matrices were used for our gene regulatory simulations.

IDENTIFYING FEEDBACK LOOPS AND FEEDFORWARD MOTIFS

Feedback and feedforward motifs appear recurrently as regulatory building blocks in transcription networks across all living organisms. These network motifs have several characteristic features (Alon, 2007), negative feedback can for example accommodate fast transcriptional responses and homeostasis, while positive feedbacks are utilized as biological switches. We went through all 1881 gene regulatory models and extracted information about their feedback and feedforward loop characteristics from their connectivity matrices. For each system we computed three autoregulatory feedback loop products $FL_1 = A_{11}$, $FL_2 = A_{22}$, $FL_3 = A_{33}$, three two-gene feedback loop products: $FL_{12} = A_{21}A_{12}$, $FL_{13} = A_{31}A_{13}$, $FL_{23} = A_{23}A_{32}$ and two three-gene feedback loop products: $FL_{123} = A_{32}A_{21}A_{13}$, $FL_{213} = A_{31}A_{12}A_{23}$. Non-zero loop products indicate that the system contains the corresponding feedback loop, and the sign of the loop product gives the sign of the feedback loop. We also computed the products for two feedforward motifs: $FFL_{32} = A_{32}(A_{31}A_{12})$,

$FFL_{31} = A_{31}(A_{32}A_{21})$. Again non-zero products indicate that the system contains the corresponding feedforward motif, a positive value corresponds to a coherent feedforward while a negative value indicates incoherent feedforward. **Figure 4** depicts the connectivity matrix and the signed digraphs of a system with a positive feedback loop as well as a system with incoherent feedforward. Spreadsheet S1 contains adjacency matrices and loop products for all 1881 motifs.

GENE REGULATORY NETWORK SIMULATIONS

The simulation were performed with the Python package `cgptoolbox` (http://github.com/jonovik/cgptoolbox), using the `sigmoidmodel` submodule, which contains an implementation of the gene regulatory network model (Equation 3) and the connectivity matrix A. A similar simulation setup is found in Gjuvsland et al. (2011) together with a discussion of gene regulation functions and the genotype-parameter map in molecular terms. We compared two different types of genotype-to-parameter maps:

- *Genotype to parameter map without pleiotropy:* biallelic genotypic variation for all three loci was introduced through the maximal production rates α_{ki}. For each Monte Carlo simulation the allelic parameter values were sampled from $U(100, 200)$.
- *Genotype to parameter map with pleiotropy:* allelic parameter values were sampled for maximal production rates α_{ki} (sampled from $U(100, 200)$), regulation thresholds θ_{lki} (sampled from $U(20, 40)$), and regulation steepnesses p_{lki} (sampled from $U(1, 10)$).

All decay rates γ_{ki} were set equal to 10. We assembled parameter sets for all 27 diploid genotypes, and for each genotypic parameter set the system of Equation 3 was integrated numerically until convergence to a stable state. The equilibrium value of y_3 was recorded as phenotype. Datasets where the system failed to converge for one or more genotypes were discarded. For each of the 1881 motifs we performed 1000 Monte Carlo simulations.

Some Monte Carlo simulations lead to very little phenotypic variation, in the sense that the span between the largest and smallest of the 27 genotypic values was small. In order to avoid artifacts arising from the numeric ODE solver tolerance, these essentially flat GP maps were discarded. Further analysis of monotonicity and variance components were only performed on GP maps where the absolute range (maximum genotypic value – minimum genotypic value) and relative range (absolute range/mean genotypic value) were both > 0.01.

RESULTS

MEASURING MONOTONICITY OF GP MAPS

In the following we present two numerical measures for quantifying monotonicity in a GP map G with N biallelic loci. The first quantifies the monotonicity for individual loci by comparing negative and positive allele substitution effects before weighting the individual loci into an overall measure. The second utilizes isotonic regression to quantify the distance between G and the closest fully monotone GP map.

Measure 1: quantifying non-monotonicity by substitution effects

We first develop a measure of monotonicity based on the effects of substituting a single allele at locus k,

$$s^1(g^{(k)}) = G(g_{1:k-1}22g_{k+1:N}) - G(g_{1:k-1}12g_{k+1:N}), \quad (4)$$
$$s^2(g^{(k)}) = G(g_{1:k-1}12g_{k+1:N}) - G(g_{1:k-1}11g_{k+1:N}),$$

while keeping the background genotype $g^{(k)} = g_{1:k+1}g_{k+1:N}$ fixed. Monotonicity as defined by Equation 2 is equivalent to $s^i(g^{(k)}) \geq 0$ for $i = 1, 2$ across all genetic backgrounds of locus k. By taking into account also the magnitude of the substitution effects we can quantify the deviation from strict monotonicity. We start with the set $S^k = \{s^i(g^{(k)})\}$ of single allele substitution effects for locus k for $i = 1, 2$ and across all genotypic backgrounds $g^{(k)}$. The total number of elements in S^k thus becomes $2 \cdot 3^{N-1}$, and we split the set into two disjoint subsets reflecting their sign; $S^k_+ = \{s^i(g^{(k)}) \in S^k | s^i(g^{(k)}) > 0\}$ and $S^k_- = \{s^i(g^{(k)}) \in S^k | s^i(g^{(k)}) < 0\}$. We compute the sum of positive substitution effects and the sum of absolute values of negative substitution effects,

$$P_k = \sum_{S^k_+} s^i(g^{(k)}), \quad (5)$$
$$N_k = \sum_{S^k_-} \left| s^i(g^{(k)}) \right|,$$

and let $T_k = P_k + N_k$ denote the overall sum of absolute substitution effects. We then define the degree to which the GP map G is monotone with respect to locus k by,

$$m_k = \frac{|P_k - N_k|}{T_k} = \frac{\left| \sum_{g^{(k)}} \left(s^1(g^{(k)}) + s^2(g^{(k)}) \right) \right|}{\sum_{g^{(k)}} \left(|s^1(g^{(k)})| + |s^2(g^{(k)})| \right)}. \quad (6)$$

The absolute value in the numerator ensures that the measure m_k is invariant with respect to the choice of indexes for the two alleles of locus k. Interchanging the numbering of the alleles leads to the mappings $s^1(g^{(k)}) \mapsto -s^2(g^{(k)})$, $s^2(g^{(k)}) \mapsto -s^1(g^{(k)})$, which leaves the value of m_k unchanged. By the triangle inequality $m_k \leq 1$. If $m_k = 1$, then G is monotonic with respect to locus k, whereas $m_k < 1$ implies that G is order-breaking w.r.t. locus k. If $m_k = 0$, then the positive substitution effects equal the negative substitution effects in magnitude and we say that G is completely order-breaking w.r.t. locus k. This measure distinguishes well between the monotone and non-monotone maps in **Figure 1**. Clearly $m_1 = m_2 = 1$ for the additive map (A) and GP maps showing partial dominance and duplicate dominance epistasis. In contrast, $m_1 = m_2 = 0$ for the maps showing pure OD and pure epistasis (A × A and D × D).

In order to quantify the overall monotonicity of the GP map G we introduce the *degree of monotonicity (m)* which is a weighted mean of all m_k, where the weights reflect the relative effect size of

the loci in terms of T_k,

$$m = \frac{\sum_{k=1}^{N} m_k T_k}{\sum_{k=1}^{N} T_k}. \quad (7)$$

As shown in **Figure 3A**, the *degree of monotonicity* is accordingly 1 for the monotone maps in **Figure 1** while it is 0 for the pure OD and pure epistasis maps. This definition of *degree of monotonicity* allows us to establish a vocabulary that is analogous to the classification of single locus dominance; i.e., a GP map is called *monotone* if $m = 1$, *(partially) non-monotone* if $m < 1$ and *purely non-monotone* if $m = 0$.

For example, the degree of monotonicity of the GP map published by Cheverud and Routman (1995), with two loci underlying 10-week body-weight (in grams) at 10 weeks in a mouse F_2 cross, may be computed as follows. After renaming the two loci (B →1, A →2) and indexing alleles to conform to our notation, the nine genotypic values (Table 1 in (Cheverud and Routman, 1995)) are $G(1111) = 31.23$, $G(1112) = 34.13$, $G(1122) = 33.82$, $G(1211) = 34.89$, $G(1212) = 35.90$, $G(1222) = 36.53$, $G(2211) = 34.12$, $G(2212) = 37.95$, and $G(2222) = 36.84$. From the line plot of this GP map (**Figure 2**, left panel) we find that the GP map is non-monotone with respect to both loci. Locus 1 shows marginal OD for the 11 genotype of locus 2 and locus 2 shows marginal OD for the 11 and 22 genotypes of locus 1. To compute the degree of monotonicity, we start with the set of single allele substitution effects for locus 1, $S^1 = \{3.66, -0.77, 1.77, 2.05, 2.71, 0.31\}$, and divide this into sets of negative $S^1_- = \{-0.77\}$ and positive effects $S^1_+ = \{3.66, 1.77, 2.05, 2.71, 0.31\}$. The sum N_1 of elements in S^1_+ is 10.50 and P_1 the sum of absolute values of elements in S^1_- is 0.77, which gives $T_1 = P_1 + N_1 = 11.27$. From Equation 6 it follows that $m_1 = 0.86$. Similarly, the sets of substitution effects for locus 2 are $S^2_- = \{-1.11, -0.31\}$ and $S^2_+ = \{3.83, 0.63, 1.01, 2.90\}$. This gives, $N_2 = 1.42$, $P_2 = 8.37$, $T_2 = 9.79$, and $m_2 = 0.71$. Inserting values for both loci into Equation 7, the degree of monotonicity (m) of this GP map is calculated to be 0.79. This value concords well with the visual observation (**Figure 2**, left panel) that it does not deviate substantially from a purely monotone map.

For random GP maps (randomly sampled genotypic values as in (Gjuvsland et al., 2011)) there is a strong positive correlation between the degree of monotonicity and the size of the additive component (V_A/V_G) (**Figure 3A**). A similar relationship was observed for three-locus random GP maps (**Figure A1A**). All GP maps in **Figure 3A** with $m < 0.1$ have $V_A/V_G < 0.1$. At the other end of the spectrum there is much more variation, for instance the most extreme completely monotone map (the duplicate dominant factors DD) has V_A/V_G as low as 0.375.

Measure 2: quantifying monotonicity by isotonic regression

This measure quantifies the monotonicity of a particular GP map G in terms of the least-squares distance to the closest monotone map. We build on the mathematical notation introduced in section "Background on monotonicity of GP maps" where Γ is the genotype space for N biallelic loci and a GP map is a function that assigns a real-valued genotypic value $G(g)$ to each genotype

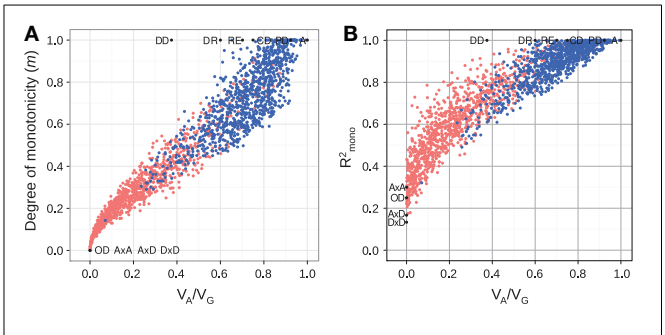


FIGURE 3 | Measures of monotonicity vs. additivity of GP maps. Scatterplots showing V_A/V_G from unweighted regression vs. **(A)** degree of monotonicity (m) and **(B)** R^2_{mono} from isotonic regression. Black dots correspond to the maps shown in **Figure 1** together with additive-by-dominance epistasis (A × D), a map with two loci showing complete dominance (CD) and two classical epistasis types from Table 1 in Phillips (1998); duplicate recessive genes (DR) and recessive epistasis (RE). Red dots show 1000 random two-locus GP maps, while blue dots show the same 1000 GP maps after rearranging genotypic values to introduce order-preservation for 1 locus [see Model and Methods in Gjuvsland et al. (2011)].

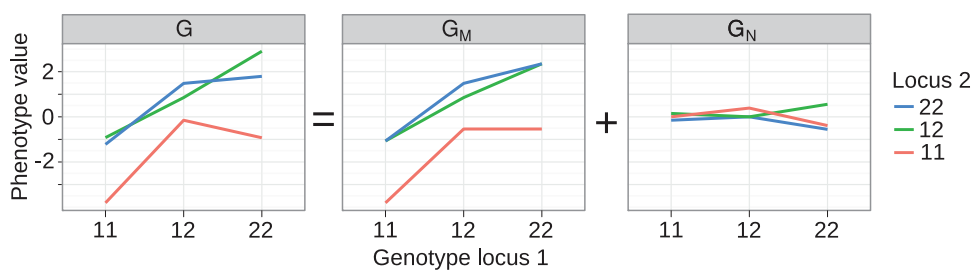


FIGURE 2 | Decomposition of genotype-phenotye map into monotone and non-monotone components. Left panel: Genotype-phenotype map G for two loci underlying 10-week body-weight at 10 weeks in a mouse F_2 cross. The GP map shown here is equivalent to the one in the original publication [see Figure 3A in Cheverud and Routman (1995)], but we have changed indexing of loci and alleles for consistency with the notation used here. The GP map G is decomposed with isotonic regression into a **(middle panel)** monotone component G_M and a (**right panel**) non-monotone component G_N.

g in Γ. For any particular GP map G, we identify the *monotone component* of G as the map G_M which minimizes the residual variance $\mathrm{var}(G - G_M)$, i.e., G_M is the monotone GP map which is closest to G in the least-squares sense. For a given G the monotone component G_M is unique (Barlow and Brunk, 1972) and can be computed numerically by isotonic regression (Leeuw et al., 2009) of G subject to the partial ordering of genotypes defined in Equation 1. Furthermore, the residual G_N is orthogonal to G_M in the sense that $\sum_{g\in\Gamma} G_M(g)G_N(g) = 0$. This allows the orthogonal decomposition,

$$G = G_M + G_N, \tag{8}$$

of a genotype-phenotype map into a *monotone component* G_M and a *non-monotone component* G_N such that $\mathrm{var}(G) = \mathrm{var}(G_M) + \mathrm{var}(G_N)$. The orthogonality property allows us to measure monotonicity of G in terms of the coefficient of determination R^2_{mono} of the isotonic regression given by the ratio $R^2_{\mathrm{mono}} = \mathrm{var}(G_M)/\mathrm{var}(G)$. In the case that G itself is monotone for all loci we have $R^2_{\mathrm{mono}} = 1$, while order-breaking for one or more loci will result in $R^2_{\mathrm{mono}} < 1$.

The isotonic regression approach can be illustrated in a straightforward way on the two-locus GP map provided by Cheverud and Routman (1995) (see text above and left panel of **Figure 2**). The partial ordering of genotypes defined by Equation 1 is illustrated in **Figure 1** (left panel). By isotone regression (Leeuw et al., 2009) on this partial genotype ordering, the original GP map is decomposed into a monotone and a non-monotone component (**Figure 2**, middle and right panels), and the coefficient of determination (R^2_{mono}) is 0.97.

Our simulation results for random GP maps show that R^2_{mono} is positively correlated to the size of the additive component (**Figure 3B** for two-locus GPs maps and **Figure A1B** for three-locus GP maps) and that for a given V_A/V_G the lower bound for R^2_{mono} is close to a straight line from (0, 0.2) to (1, 1). However, due to the search for the closest monotone GP map, R^2_{mono} will not become zero even for purely overdominant or purely epistatic maps. As shown in **Figure A2**, the two monotonicity measures are highly correlated.

An R package for studying monotonicity in GP maps

We developed an R package `gpmap` for studying functional properties of GP maps. The package takes GP maps in the form of vectors of genotypic values as input, and provides functions for (i) determining whether the map is order-breaking or order-preserving w.r.t. any given locus, (ii) the degree of monotonicity m, (iii) R^2_{mono} using isotonic regression from the `isotone` package (Leeuw et al., 2009), and (iv) plots of the original and decomposed GP maps. Code example 1 (**Box 1**) below illustrates the usage and functionality of the `gpmap` package. The package is available from CRAN http://cran.r-project.org/package=gpmap under GPLv3.

MONOTONICITY IN GP MAPS ARISING FROM GENE REGULATORY NETWORKS

To search for generic relationships between monotonicity and regulatory network structure, we used the above measures of monotonicity to characterize GP maps emerging from the gene regulatory network models (see Models and Methods). Based on earlier results (Gjuvsland et al., 2007, 2011; Wang et al., 2013) we hypothesized that incoherent feed forward (**Figure 4**, right panel) or positive feedback (**Figure 4**, left panel) would be necessary in order to obtain highly order-breaking GP maps, and we characterized all 1881 networks in terms of these two properties. **Table 1** shows the number of motifs falling into the resulting four categories. We summarized the number of Monte Carlo simulations where all genotypic parameter sets gave convergence to a stable steady state, and where the resulting GP maps were not essentially flat (see Models and Methods for details). Motifs with less than 100 usable GP maps were discarded from further analysis. For the genotype-to-parameter maps without pleiotropy (in the sense

Box 1 | Code example 1.

Code example for quantifying and visualizing monotonicity for the two-locus GP map published in [14] using the R package `gpmap`.

```
> library(gpmap) #load package
> data(GPmaps) #load dataset
> gp <- mouseweight #GP map from reference
  [14]
>
> ## Tabulate genotypic values
> cbind(gp$genotype,gp$values)
>
> ## Plot the GP map
> plot(gp)
>
> ## Compute degree of monotonicity
> gp <- degree_of_monotonicity(gp)
> gp$degree.monotonicity.locus
> print(gp)
>
> ## Quantify monotonicity by isotonic
  regression
> gp <- decompose_monotone(gp)
> print(gp)
>
> ## Plot decomposed GP map
> plot(gp,decomposed=TRUE)
```

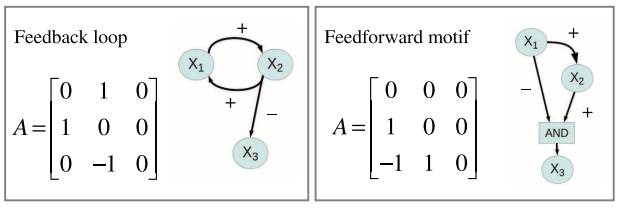


FIGURE 4 | Connectivity matrices and signed directed graphs. Connectivity matrix A and the corresponding signed directed graph for two of the 1881 systems in the simulation study. The **left panel** depicts the connectivity matrix and the signed digraph of a system with a positive feedback loop between X_1 and X_2 while the **right panel** shows a system with incoherent feedforward from X_1 to X_3.

Table 1 | Frequencies (proportion of row total in parenthesis) of incoherent feedforward and positive feedback loops in subsets of the 1881 studied motifs.

Dataset	Number of motifs	Motifs containing			
		Incoh. feedforward		No incoh. feedforward	
		Positive feedback	No positive feedback	Positive feedback	No positive feedback
All motifs	1881	287 (0.153)	48 (0.026)	1294 (0.688)	252 (0.134)
GENOTYPE-TO-PARAMETER MAP WITHOUT PLEIOTROPY					
Discarded motifs	868	152 (0.175)	0	715 (0.824)	1 (0.001)
Analyzed motifs	1013	135 (0.133)	48 (0.047)	579 (0.571)	251 (0.248)
GENOTYPE-TO-PARAMETER MAP WITH PLEIOTROPY					
Discarded motifs	791	124 (0.157)	0	667 (0.84)	0
Analyzed motifs	1090	163 (0.149)	48 (0.044)	627 (0.575)	252 (0.231)

that genetic variation at one locus influences only a single parameter, see Model and Methods) 868 motifs were discarded, while for the genotype-to-parameter map with pleiotropy (genetic variation at one locus influences three parameters) 791 motifs were discarded. All (but one) discarded motifs contained at least one positive feedback loop (**Table 1**). A plausible explanation for this is that many motifs with positive feedback loops have a stable steady state at, or very close to 0 for one or more state variables regardless of parameter values, and this leads to essentially flat GP maps.

The introduction of pleiotropy in the genotype to parameter map has a marked effect on the monotonicity characteristics of the associated GP map (**Figure 5**). When genetic variation at a locus X_i affects only its maximal production rate the GP maps come out as highly monotone (**Figure 5A**), with a large majority being fully monotone or order-breaking for just a single locus. When genetic variation at locus X_i affects the threshold and steepness of the dose-response curve in addition to the maximal production rate (pleiotropy in the genotype-to-parameter map), the majority of GP maps still show order-breaking either for no loci or just one locus (**Figure 5B**). But a considerable number of GP maps are in this case order-breaking for two or three loci. Furthermore, by dividing the motifs into the four groups given in **Table 1** it is evident that the regulatory anatomy of a network determines its predisposition for non-monotonicity in its associated GP map. Presence of incoherent feedforward or positive feedback loops appears to be prerequisites for the majority of the observed non-monotonic GP maps.

The class of motifs lacking both incoherent feedforward and positive feedback contains very few order-breaking GP maps, and with no pleiotropy in the genotype-to-parameter map we observe only fully order-preserving GP maps for this class (cyan in **Figure 5A**). In the Appendix we generalize this to an arbitrary number of nodes and formally prove that without pleiotropy in the genotype-to-parameter map, the presence of incoherent feedforward or positive feedback is indeed a necessary condition for non-monotone GP maps to arise from networks with monotone gene regulation functions.

The introduction of pleiotropy in the genotype-to-parameter map increases the frequency of order-breaking GP maps substantially (**Figure 5B**). Motifs lacking both incoherent feedforward and positive feedback may in this case lead to GP maps that are order-breaking for one or two loci, but never for all three loci. Using isotonic regression to quantify the overall monotonicity of the GP maps reinforces the finding that incoherent feedforward and positive feedback predispose for non-monotonicity (**Figure 6**). **Figure 6** also shows that for all classes of motifs the majority of GP maps are fully monotone, while the most non-monotone GP maps (lowest R^2_{mono}values) are observed for motifs with positive feedback. The differences between classes

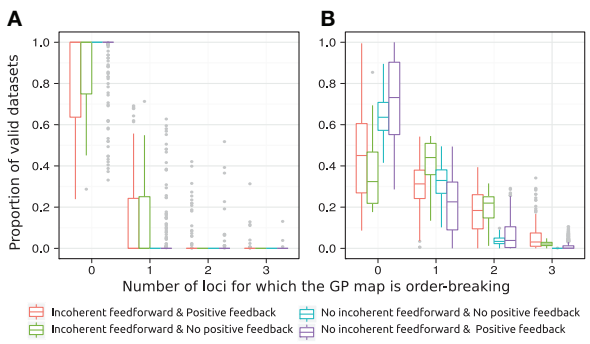


FIGURE 5 | Order-breaking in motifs containing a single feedforward loop. Summary of order-breaking for all motifs for which at least 100 (out of 1000) Monte Carlo simulations lead to GP maps with non-negligible variation (see Models and Methods section "Gene regulatory network simulations," for detailed criteria). Results are shown for 1013 motifs with a genotype-to-parameter map without pleiotropy **(A)** and 1090 motifs with a genotype-to-parameter map with pleiotropy **(B)**. Colors indicate classes of motifs based on the presence/absence of incoherent feedforward and positive feedback loops, see **Table 1** for the number of motifs in each class. A single boxplot summarizes, for all motifs in the given class, the proportion of the GP maps (y-axis) that are order-breaking with respect to a given number of loci (x-axis). For example, consider the red box at $x = 0$ in panel **(A)**. This boxplot contains results for motifs with both incoherent feedforward and positive feedback and from **Table 1** we find that the red boxplot summarizes results for 135 motifs. From the y-axis we find that at least half (box median at $y = 1$) of these 135 motifs result in only monotone GP maps, while for the most extreme (end of whisker) of the 135 motifs only 25% of the GP maps are monotone. Similarly, the cyan box is compressed into a line at $x = 0$, $y = 1$ indicating that all 251 motifs that lack both incoherent feedforward and positive feedback result in only monotone GP maps.

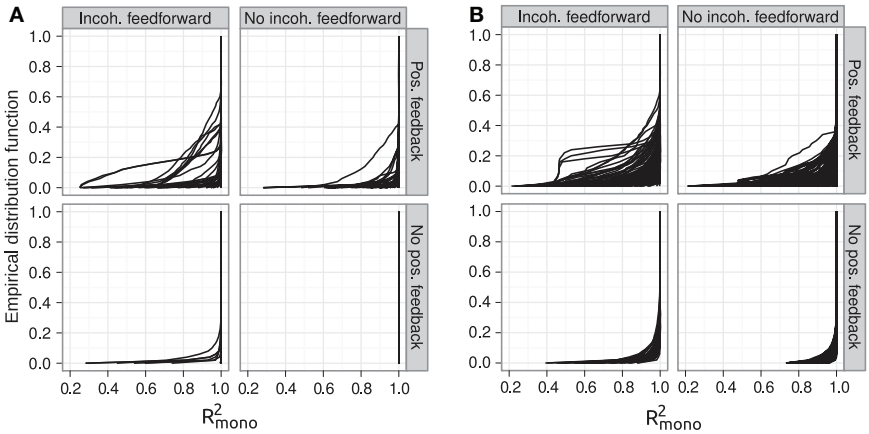


FIGURE 6 | Empirical distribution functions for R^2_{mono}. Summary of R^2_{mono} values from isotone regression for all motifs for which at least 100 (out of 1000) Monte Carlo simulations lead to GP maps with non-negligible phenotypic variation (see Models and Methods section "Gene regulatory network simulations," for detailed criteria). Results are shown for 1013 motifs with a genotype-to-parameter map without pleiotropy **(A)** and 1090 motifs with a genotype-to-parameter map with pleiotropy **(B)**. Each panel is divided into 4 subplots containing classes of motifs based on the presence/absence of incoherent feedforward and positive feedback loops, see **Table 1** for the number of motifs in each class. Each curve shows, for a single motif, the empirical distribution function value (y-axis) of R^2_{mono} for all GP maps (x-axis).

of motifs are also evident when inspecting the additivity of GP maps (**Figure A3**), but since monotone GP maps can still be non-additive, the patterns are much more blurred than for monotonicity.

DISCUSSION

Fisher's (1918) regression on gene content and the concepts derived from this, such as additive effects and dominance deviation, provide the theoretical basis for most of quantitative genetics (Falconer and Mackay, 1996; Lynch and Walsh, 1998). By regressing on gene content, including the extensions by Cockerham (1954), the genotype-phenotype map is decomposed into additive, dominant, and epistatic components. The use of gene content or the number (0, 1, or 2) of alleles with a particular index in a genotype implies the same partial ordering of genotype space as defined in Equation 1. Thus, our proposed definition of monotonicity of GP maps, and in particular the use of isotonic regression to quantify monotonicity, may be viewed as a relaxation of the linearity assumption underlying current quantitative genetics theory. In this perspective the positive correlation between monotonicity and additivity (**Figure 3**) is expected.

We have addressed GP maps with 2 and 3 loci as we considered an in-depth study of the properties of GP maps with higher number of loci to be outside the scope of this study. Some general observations can be made, though. Since m is a weighted average, the m_k of major loci (i.e., for which T_k is large relative to $\sum T_k$) will tend to dominate. For instance, in a case with a single major locus showing monotone gene action and several minor loci showing order-breaking, the GP map will overall be close to monotone (m close to 1). Conversely, order-preservation in a number of minor loci would have little influence on m if major loci have strongly non-monotone gene action. Isotonic regression gives an overall measure of monotonicity of a GP map, but provides no locus-specific measures corresponding to m_k. Similar to the case for m, the gene action of major loci will have high influence on the value of R^2_{mono}.

The observation that monotonicity is an important property of GP maps is in principle not new. For a single locus, non-monotone gene action appears in the form of over- or under-dominance, while complete and partial dominance as well as additivity exemplify monotone gene action. Weinreich et al. (2005) distinguished between *sign epistasis* and *magnitude epistasis* and showed that sign epistasis limits the number of mutational trajectories to higher fitness. As sign epistasis reflects a non-monotone GP relationship and magnitude epistasis reflects a monotone one, this insight concords with our results. A similar distinction has been proposed (Wang et al., 2010) for statistical interactions where *removable interactions* are those that can be removed by a monotone transformation of the phenotype scale, while non-monotonicity in the GP map leads to *essential interactions*. Wu et al. (2009) developed a method to screen for and test the significance of essential interaction in genome-wide association studies. Isotonic regression has also recently been applied to link genotype and phenotype data (Beerenwinkel et al., 2011; Luss et al., 2012). Our treatment of monotonicity is more general than these earlier works in three major ways. First, we deal with monotonicity of the GP map as a whole rather than either intra-locus (dominance vs. overdominance) or inter-locus (magnitude vs. sign epistasis and removable vs. essential interactions). Second, where the earlier treatments have focused on classifying the type of gene action, we make use of quantitative measures of monotonicity. Third, our approach combining the concept of monotonicity with cGP models opens a direct link between genetics and the theory of dynamical systems in the wide sense.

Monotonicity is a property of the GP map separate from the allele frequencies, making it a physiological (Cheverud and Routman, 1995) or functional (Hansen and Wagner, 2001)

descriptor rather than a statistical one. The distinction between physiological and statistical epistasis has lead to much debate (Phillips, 2008). Zeng et al. (2005) argued the distinction was unnecessary and potentially misleading. Although their arguments around orthogonality and variance components are valid, our results demonstrate very clearly that describing the properties of the GP map without reference to any particular study population is essential if we want to connect quantitative genetics with regulatory biology.

It is clear from our results that positive feedback and incoherent feedforward promote non-monotonicity. The clear-cut differences in monotonicity between different classes of regulatory networks, combined with the strong correlation between monotonicity and additivity of GP maps, appear therefore to explain the findings that regulatory systems with positive feedback give considerably more statistical epistasis than those without (Gjuvsland et al., 2007; Wang et al., 2013). Even though both incoherent feedforward and positive feedback predispose for non-monotone GP maps, the underlying mechanisms are different for the two regulatory motifs. In the case of incoherent feedforward the sum of direct and indirect effects may result in a non-monotone dose-response relationship (Kaplan et al., 2008). That positive feedback loops can give non-monotonicity is intuitively less clear, but in the Appendix we show both results analytically. Positive feedback predisposes for multiple steady states, and order-breaking might also emerge from different genotypes corresponding to different states. It should be noted, however that positive feedback is only a necessary condition for multistationarity (Plahte et al., 1995), and a positive loop in the connectivity matrix A of a system is not necessarily active at any point during the time course of the system.

Without any restrictions on the connectivity of a three-gene system there are $3^9 = 19,683$ possible distinct networks. The main restriction we imposed (see Models and Methods for details) was a maximum of two regulators per gene, which allowed us to use Boolean gene regulation functions already established in the sigmoid formalism (Plahte et al., 1998). Other model formalisms allowing an arbitrary number of regulators are also available (Wagner, 1994, 1996; Siegal and Bergman, 2002) and could be extended to diploid forms and used in later studies.

Although this study has focused on gene regulatory networks, the concept of monotone gene action applies to the propagation of genetic variation across the whole physiological hierarchy. One may therefore systematically use the concepts and methods presented here to study the order-preserving and order-breaking properties of genotype-phenotype mappings that are associated with any regulatory structure amenable for mathematical modeling. Through this it will be possible to make a wide-ranging survey of which regulatory anatomies promote monotonicity and which promote non-monotonicity. We foresee that this classification may become instrumental for predicting how phenotypic effects of genetic variation propagate across generations in sexually reproducing populations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2013.00216/abstract

Spreadsheet S1 | Excel spreadsheet with connectivity matrices and loop products for all 1881 gene regulatory networks.

REFERENCES

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461. doi: 10.1038/nrg2102

Alvarez-Castro, J. M., and Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in QTL analysis. *Genetics* 176, 1151–1167. doi: 10.1534/genetics.106.067348

Barlow, R. E., and Brunk, H. D. (1972). The isotonic regression problem and its dual. *J. the Am. Stat. Assoc.* 67, 140–147. doi: 10.1080/01621459.1972.10481216

Beerenwinkel, N., Knupfer, P., and Tresch, A. (2011). Learning monotonic genotype-phenotype maps. *Stat. Appl. Genet. Mol. Biol.* 10:3. doi: 10.2202/1544-6115.1603

Cheverud, J. M., and Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics* 139, 1455–1461.

Cheverud, J. M., and Routman, E. J. (1996). Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution* 50, 1042–1051. doi: 10.2307/2410645

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variances for analysis of covariances among relatives when epistasis is present. *Genetics* 39, 859–882.

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103. doi: 10.1089/10665270252833208

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics.* Harlow: Longman Group.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52, 399–433. doi: 10.1017/S0080456800012163

Gjuvsland, A. B., Hayes, B. J., Omholt, S. W., and Carlborg, Ö. (2007). Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* 175, 411–420. doi: 10.1534/genetics.106.058859

Gjuvsland, A. B., Vik, J. O., Beard, D. A., Hunter, P. J., and Omholt, S. W. (2013). Bridging the genotype-phenotype gap: what does it take? *J. Physiol.* 591, 2055–2066. doi: 10.1113/jphysiol.2012.248864

Gjuvsland, A. B., Vik, J. O., Woolliams, J. A., and Omholt, S. W. (2011). Order-preserving principles underlying genotype-phenotype maps ensure high additive proportions of genetic variance. *J. Evol. Biol.* 24, 2269–2279. doi: 10.1111/j.1420-9101.2011.02358.x

Hansen, T. F., and Wagner, G. P. (2001). Modeling genetic architecture: a multilinear theory of gene interaction. *Theor. Popul. Biol.* 59, 61–86. doi: 10.1006/tpbi.2000.1508

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* 4:e1000008. doi: 10.1371/journal.pgen.1000008

Jaeger, J., Irons, D., and Monk, N. (2012). The inheritance of process: a dynamical systems approach. *J. Exp. Zool. B Mol. Dev. Evol.* 318, 591–612. doi: 10.1002/jez.b.22468

Kaplan, S., Bren, A., Dekel, E., and Alon, U. (2008). The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol. Syst. Biol.* 4, 203. doi: 10.1038/msb.2008.43

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 9, 770–780. doi: 10.1038/nrm2503

Latchman, D. (2005). *Gene Regulation: a Eukaryotic Perspective.* New York, NY: Taylor and Francis.

Le Rouzic, A., and Alvarez-Castro, J. M. (2008). Estimation of genetic effects and genotype-phenotype maps. *Evol. Bioinformatics* 4, 225–235. doi: 10.4137/EBO.S756. Available online at: http://www.la-press.com/estimation-of-genetic-effects-and-genotype-phenotype-maps-article-a887

Leeuw, J. D., Hornik, K., and Mair, P. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *J. Stat. Softw.* 32, 1–24.

Luss, R., Rosset, S., and Shahar, M. (2012). Efficient regularized isotonic regression with application

to gene–gene interaction search. *Ann. Appl. Stat.* 6, 253–283. doi: 10.1214/11-AOAS504

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer Associates.

Mestl, T., Plahte, E., and Omholt, S. W. (1995). A mathematical framework for describing and analysing gene regulatory networks. *J. Theor. Biol.* 176, 291–300. doi: 10.1006/jtbi.1995.0199

Moore, A. (2012). Towards the new evolutionary synthesis: gene regulatory networks as information integrators. *Bioessays* 34, 87–87. doi: 10.1002/bies.201290000

Neale, B. M., Ferreira, M. A. R., Medland, S. E., and Posthuma, D. (eds.). (2008). *Statistical Genetics: Gene Mapping through Linkage and Association*. New York, NY: Taylor and Francis Group.

Omholt, S. W., Plahte, E., Øyehaug, L., and Xiang, K. F. (2000). Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* 155, 969–980.

Phillips, P. C. (1998). The language of gene interaction. *Genetics* 149, 1167–1171.

Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9, 855–867. doi: 10.1038/nrg2452

Plahte, E., Mestl, T., and Omholt, S. W. (1995). Feedback loops, stability and multistationarity in dynamical systems. *J. Biol. Syst.* 3, 409–413. doi: 10.1142/S0218339095000381

Plahte, E., Mestl, T., and Omholt, S. W. (1998). A methodological basis for description and analysis of systems with complex switch-like interactions. *J. Math. Biol.* 36, 321–348. doi: 10.1007/s002850050103

Schlitt, T., and Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics* 8(Suppl. 6):S9. doi: 10.1186/1471-2105-8-S6-S9

Siegal, M. L., and Bergman, A. (2002). Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10528–10532. doi: 10.1073/pnas.102303999

Wagner, A. (1994). Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. U.S.A.* 91, 4387–4391. doi: 10.1073/pnas.91.10.4387

Wagner, A. (1996). Does evolutionary plasticity evolve? *Evolution* 50, 1008–1023. doi: 10.2307/2410642

Wang, X., Elston, R. C., and Zhu, X. (2010). The meaning of interaction. *Hum. Hered.* 70, 269–277. doi: 10.1159/000321967

Wang, Y., Vik, J. O., Omholt, S. W., and Gjuvsland, A. B. (2013). Effect of regulatory architecture on broad versus narrow sense heritability. *PLoS Comput. Biol.* 9:e1003053. doi: 10.1371/journal.pcbi.1003053

Weinreich, D. M., Watson, R. A., and Chao, L. (2005). Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59, 1165–1174. doi: 10.1554/04-272

Wu, C., Zhang, H., Liu, X., Dewan, A., Dubrow, R., Ying, Z., et al. (2009). Detecting essential and removable interactions in genome-wide association studies. *Stat. Interface* 2, 161–170. doi: 10.4310/SII.2009.v2.n2.a6

Zeng, Z. B., Wang, T., and Zou, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* 169, 1711–1725. doi: 10.1534/genetics.104.035857

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 June 2013; accepted: 07 October 2013; published online: 07 November 2013.

*Citation: Gjuvsland AB, Wang Y, Plahte E and Omholt SW (2013) Monotonicity is a key feature of genotype-phenotype maps. Front. Genet. **4**:216. doi: 10.3389/fgene.2013.00216*

This article was submitted to Genetic Architecture, a section of the journal Frontiers in Genetics.

APPENDIX

In this appendix we complement the simulation studies in the main text with some analytic results for GP maps emerging from ODE models of gene regulatory networks. We study a generalization of the gene network model in Equation (3) with an arbitrary number of loci and monotone gene regulation functions, but restrict the analysis to genotype-parameter maps without pleiotropy. In particular, we show that (i) if there are no positive feedback loops and no incoherent feedforward loops in the network, the resulting GP maps are always monotone, (ii) a positive feedback loop or an incoherent feedforward loop may lead to non-monotone GP maps. The results hold for phenotypes given as the stable concentration of the product of one of the genes, and under certain restrictions also for phenotypes given as a function of one or several stable gene product concentrations that is monotonic with respect to each of its arguments.

GENE NETWORK MODEL

We consider a dynamic system consisting of n mutually interacting diploid loci X_j, $j \in N = \{1, \ldots, n\}$, regulating each other's expression. The time dependent output of X_j is denoted z_j, and we define $z = [z_1, z_2, \ldots, z_n]$. It goes without saying that z_j in general depends on the genotypes of all the genes even though we will not always state this explicitly.

For a given genotype $g = g_j g^{(j)} = a_j b_j g^{(j)}$, where $g_j \in \{11, 12, 22\}$ denotes the genotype and $a_j, b_j \in 1, 2$ denote the indexes of the two alleles of locus X_j, the equations of motion for X_j are

$$\begin{aligned} \dot{z}_j^1 &= \alpha_j^{a_j} r_j^{a_j}(z) - \gamma_j^{a_j} z_j^1, \\ \dot{z}_j^2 &= \alpha_j^{b_j} r_j^{b_j}(z) - \gamma_j^{b_j} z_j^2, \\ z_j &= z_j^1 + z_j^2, \end{aligned} \tag{A1}$$

where z_j^1 and z_j^2 are the time-dependent outputs of the two homologous copies of X_j. The two allele rate functions $r_j^1(z)$ and $r_j^2(z)$ have range $[0, 1]$ so that α_j^1 and α_j^2 represent the maximum production rates of the two alleles. We assume that all dose-response functions in Equation (A1) are differentiable and monotonic with respect to each of its arguments, and that for each j, k, the signs of $\partial r_j^1/\partial x_k$ and $\partial r_j^2/\partial x_k$ in the stable point x are equal. This model generalizes Eq. (3) to an arbitrary number of loci and a broader class of gene regulation functions.

In the following we are only concerned with the steady states of Equation (A1), and assume for simplicity that they have just a single stable equilibrium point. Solving the equilibrium conditions of Equation (A1) with respect to z_j^1 and z_j^2 and adding gives

$$f_j(x) = \mu_j^{a_j} r_i^{a_j}(x) + \mu_j^{b_j} r_j^{b_j}(x) - x_j = 0, \quad j \in N, \tag{A2}$$

where $x = [x_1, \ldots, x_n]$ is the stable point, $\mu_j^{a_i} = \alpha_j^{a_j}/\gamma_j^{a_j}$ and $\mu_j^{b_j} = \alpha_j^{b_j}/\gamma_j^{b_j}$. Since our definition of monotonicity of GP maps does not depend on the numbering of alleles, we will without loss of generality assume $\mu_j^1 \leq \mu_j^2$ for all j.

The network architecture can be read out from the structure of the system's Jacobian matrix in the stable state x. We define the elements of the Jacobian J for the set of functions f_j defined in Equation (A2) by

$$J_{jk} = J_{jk}(g) = \frac{\partial f_j(x)}{\partial x_k}, \quad j, k \in N. \tag{A3}$$

To the Jacobian J it is customary to assign a signed directed graph $\mathcal{G}$ in which each locus X_k is represented by a node X_k, and in which there is an arc from X_j to X_k if and only if $J_{kj} \neq 0$, its sign given by the sign of J_{kj}. A chain from X_j to X_k is a set of arcs in $\mathcal{G}$ leading from X_j to X_k in which all intermediate nodes are visited only once. The sign of a chain is equal to the product of the signs of the J_{ij} corresponding to the arcs in the chain. If there is a chain from X_i to X_j and also a chain from X_j to X_i through a disjoint set of nodes, the two chains constitute a proper feedback loop (FBL). To each FBL is associated a loop product L which is the product of the Jacobian elements corresponding to all the arcs in the loop. The sign of the loop is given by the sign of L. Two chains from X_j to X_i, $i \neq j$, with only the endpoint nodes in common, constitute a feedforward loop (FFL). If the two chains have opposite signs, the FFL is incoherent (IFFL), otherwise it is coherent (CFFL).

The system's phenotype could be any scalar quantity defined by its equilibrium value x. In the following we assume the genotype-phenotype map $G(g) = x_q(g)$, $q \in N$, for a given and fixed q, and investigate the monotonicity properties of $G(g_k g^{(k)})$ with respect to genetic variation in any locus X_k for different backgrounds $g^{(k)}$. In the following sections we analyse the causes of order-breaking in G in the restricted case in which there is only genetic variation in μ_k^1 and μ_k^2, not in the shape of the dose-response functions r_k^1 and r_k^2, implying $r_k^1(x) = r_k^2(x) = r_k(x)$. This is what we mean by a genotype-to-parameter map without pleoitropy.

In the next sections we prove the following result:

Proposition 1. *Assume all rate functions in Equation* (A1) *are monotonic and that G is the mapping from g to x_q for some fixed q so that $x_q(g)$ is the phenotype. If there is no feedback loop (FBL) and no feedforward loop (FFL) anywhere in the network corresponding to the system Equation* (A1)*, then necessarily $m_k = 1$ for all k. If the system contains either a single FFL or a single FBL, then G may be non-monotone for some x_k if the FFL is positive or the FBL is incoherent, but if the FBL is negative or the FFL is coherent, no order breaking can occur for any x_k.*

At the end of this note we show that under some reasonable conditions this result is also valid for more general phenotypes depending on more than one x_q.

NETWORKS WITHOUT LOOPS

We first consider networks containing no feedforward loop and no feedback loop. In these networks there is at most one chain from one node to another, and of course no autoregulatory loops. If there is a chain from X_j to X_k, there is no chain from X_k to X_j. Any node is either unregulated (constitutively expressed) or regulated by one or several other nodes.

We first prove a useful lemma.

Lemma 1. *If $x_l(11g^{(j)}) \leq x_l(12g^{(j)}) \leq x_l(22g^{(j)})$ for any j and l and there is an arc $X_l \to X_m$ with positive sign and no other chain from $X_l \to X_m$, then also $x_m(11g^{(j)}) \leq x_m(12g^{(j)}) \leq x_m(22g^{(j)})$. If the sign of the arc is negative, then $x_m(11g^{(j)}) \geq x_m(12g^{(j)}) \geq x_m(22g^{(j)})$.*

Proof. Suppressing the explicit dependence on other genes that are not affected by genetic variation in X_j, we have

$$\begin{aligned} x_m(11g^{(j)}) &= 2\mu_m^1 r_m(x_l(11g^{(j)})), \\ x_m(12g^{(j)}) &= (\mu_m^1 + \mu_m^2) r_m(x_l(12g^{(j)})), \\ x_m(22g^{(j)}) &= 2\mu_m^2 r_m(x_l(22g^{(j)})). \end{aligned} \tag{A4}$$

Now, r_m is monotonic by assumption. If it is monotonically increasing,

$$\begin{aligned} x_m(12g^{(j)}) &\geq (\mu_m^1 + \mu_m^2) r_m(x_l(11g^{(j)})) \geq x_m(11g^{(j)}), \\ x_m(22g^{(j)}) &\geq 2\mu_m^2 r_m(x_l(12g^{(j)})) \geq x_m(12g^{(j)}), \end{aligned} \tag{A5}$$

from which the assertion follows. If r_m is monotonically decreasing, we find the same relations with the inequality signs reversed. □

If there is no chain from X_j to X_q, genetic variation in X_j will not be reflected in G, i.e. $G(11g^{(j)}) = G(12g^{(j)}) = G(22g^{(j)})$, and by definition does not give order-breaking. Then assume X_j is upstream relative to X_q and that the chain from X_j to X_q is positive. We first let X_j be an unregulated node with no predecessor. Then

$$\begin{aligned} x_j(11g^{(j)}) &= 2\mu_j^1, \\ x_j(12g^{(j)}) &= \mu_j^1 + \mu_j^2, \\ x_j(22g^{(j)}) &= 2\mu_j^2, \end{aligned} \tag{A6}$$

because $r_j^1 = r_j^2 = 1$. From this it follows that $x_j(11g^{(j)}) \leq x_j(12g^{(j)}) \leq x_j(22g^{(j)})$.

Repeated use of Lemma 1 leads eventually to $x_q(11g^{(j)}) \leq x_q(12g^{(j)}) \leq x_q(22g^{(j)})$, irrespective of the genotypic background of X_j. If the chain from X_j to X_q is negative, the argument goes in the same way, but then $x_q(11g^{(j)}) \geq x_q(12g^{(j)}) \geq x_q(22g^{(j)})$. The above argument can be carried out in the same way if X_j is not top-stream. It follows that in a network without FFBs and FFLs and where genetic variation is restricted to μ_k^1 and μ_k^2, the genotype-phenotype map $G(g) = x_q(g)$ cannot be order-breaking.

NETWORKS WITH A FEEDBACK LOOP

In this section we investigate the effects of feedback loops on the degree of monotonicity. Assuming monotonic dose-response functions and non-pleiotropic genetic variation, we show that a positive feedback loop may lead to order breaking, while negative feedback loops never do. We consider a network in which there is no FFL and a single FBL with X_q as one of its members and X_k is upstream of the loop.

Lemma 2. *Consider a network with n nodes for which all dose-response functions are monotonic and there is only genetic variation in μ_k^1 and μ_k^2. Assume there is a chain from X_k to X_1, that X_1, but not X_k, is member of a FBL with m nodes, and that there is no other FBL and no FFL in the system. If X_q is in the loop, let the loop be $X_1 \to X_2 \to \ldots \to X_q \to \ldots \to X_m \to X_1$. If the FBL is positive, there may be order-breaking in X_q due to genetic variation in X_k, but no order-breaking can occur if the loop is negative. If X_q is downstream of the loop, the same result applies.*

Proof. With a single FBL and no FFL there is at most one directed path from any node X_i to any other node X_j, and if there is a path from X_i to X_j, there is no return path from X_j to X_i if either X_i or X_j is not part of the FBL. We first consider the dependence of x_1 on x_k. The direct regulators of node X_1 are X_m and X_l, the latter being the last but one node in the chain from X_k to X_1. In Plahte et al. (2013) we introduced *the propagation functions* $x_j = p_{jk}(x_k)$ which express the effect on x_j of genetic variation in X_k. An important property of p_{jk} is that it can be derived from all the equilibrium conditions Equation (A2) except the equation for f_k. This implies that the effects on X_j of genotypic variation in X_k are only expressed in terms of the variations in x_k, while the parameters expressing the genotype of X_k do not enter into the function p_{jk} .

We then have $x_l = p_{lk}(x_k)$ and $x_m = p_{m1}(x_1)$. To make it easier to use the results in Plahte et al. (2013) we rewrite the equilibrium condition Equation (A2) as

$$R_j(x) - \gamma_j x_j = 0, \tag{A7}$$

where $\gamma_j > 0$. In the following, the Jacobian refers to this set of equations, which has the same root and the same functional dependencies between the variables as the original set. The signs of the partial derivatives of R_j are the same as for $r_j^{a_j}$ and $r_j^{b_j}$. The equilibrium condition for X_1 is then

$$\gamma_1 x_1 = R_1(p_{lk}(x_k), p_{m1}(x_1))). \tag{A8}$$

This equation defines x_1 as a function of x_k in an open domain around the equilibrium point and with a derivative that can be computed by implicit differentiation, i.e.

$$\gamma_1 \frac{\mathrm{d}x_1}{\mathrm{d}x_k} = \frac{\partial R_1}{\partial x_l} q_{lk} + \frac{\partial R_1}{\partial x_m} q_{m1} \frac{\mathrm{d}x_1}{\mathrm{d}x_k}, \tag{A9}$$

where $q_{ij} = p'_{ij}$ is the derivative of p_{ij} for all i, j.

From Lemma 1 it follows that there is no order breaking in X_l, in other words, q_{lk} has a fixed sign. Consider then q_{m1}. There is just a single chain from X_1 to X_m, and Equation (13) in Plahte et al. (2013) gives

$$q_{m1}(x_1) = (-1)^{m-1} \frac{D_{VV} C_U}{D^{(11)}}. \tag{A10}$$

Here U is the set of nodes in this chain, C_U is its chain product, i.e. the product of the Jacobian elements corresponding to the arcs in the chain, $V = N \setminus U$, $D^{(11)}$ is the subdeterminant of J with row 1 and column 1 deleted, and D_{VV} is the subdeterminant

of J composed of the rows and columns V. Because there is no feedback loop among the nodes represented in $D^{(11)}$ and D_{VV}, only the diagonal degradation terms contribute to these two determinants. Hence $D^{(11)} = (-1)^{n-1}\prod_{i\neq 1}\gamma_i$. Similarly, $D_{VV} = (-1)^{n-m}\prod_{i\in V}\gamma_i$, giving $q_{ml} = \gamma_1 C_U/\Gamma_U$, where $\Gamma_U = \prod_{i\in U}\gamma_i$. Finally, we note that $P = (\partial R_1/\partial x_m)C_U$ is the loop product of the loop.

Solving Equation (A9) with respect to $\mathrm{d}x_1/\mathrm{d}x_k$ and using all these expressions lead to

$$\gamma_1\frac{\mathrm{d}x_1}{\mathrm{d}x_k} = \frac{\Gamma_U}{\Gamma_U - P}\frac{\partial x_1}{\partial x_l}q_{1k}. \tag{A11}$$

The sign of $\partial x_1/\partial x_l$ is independent of the genotype of X_k and the sign of q_{1k} is fixed. Genotypic variation in X_k may change the magnitude of P, but its sign is fixed because all Jacobi elements have fixed sign independent of the system parameters. Thus, genotypic variation in X_k does not alter the sign of $\mathrm{d}x_1/\mathrm{d}x_k$ if the loop is negative ($P < 0$), while for a positive loop the sign of $\Gamma_U - P$ may switch. In the latter case, an increase in x_k due to genetic variation in X_k may increase x_1 in some cases and decrease it in others, leading to order breaking. As there is only a single chain from X_1 to X_q, no order breaking in X_1 implies no order breaking in X_q, while order breaking in X_1 may propagate to X_q. The same result follows if X_q is downstream a node in the loop because order breaking in this node may propagate to X_q. □

FEEDFORWARD LOOPS (FFLS)

A feedforward loop (FFL) is a motif in the network in which there are two different chains C_1 and C_2 from one particular node to another particular node. To each chain C_i is associated a chain product P_i defined as the product of the Jacobian elements corresponding to the arcs in C_i. If P_1 and P_2 have equal signs, the FFL is coherent, otherwise it is incoherent.

In a network with a single feedforward loop and no feedback loops we now investigate the effect on $G(g) = x_q(x_k(g))$ of genetic variation in X_k for varying background $g^{(k)}$. Our starting point is again Equation (A7). We first let X_k and X_q be the initial and terminal nodes in the FFL. The two chains C_1 and C_2 leading from X_k to X_q comprise ρ_1 and ρ_2 nodes including X_k and X_q, respectively. Let the set of nodes in C_1 and C_2 be X_{U_1} and X_{U_2}, respectively, where U_1 and U_2 are the corresponding subsets of N, and let V_1 and V_2 be their complements.

Roughly speaking, the derivative of the propagation function $p_{qk}(x_k)$ can be expressed as a sum of terms, each term corresponding to one of the chains leading from X_k to X_q (Plahte et al., 2013). To the chain C_i is assigned the chain weight w_i given by

$$w_i = (-1)^{\rho_i - 1}\frac{D_{V_iV_i}}{D^{(kk)}}, \quad i = 1, 2, \tag{A12}$$

where $D_{V_iV_i}$ is the Jacobian subdeterminant for the nodes not included in C_i, and $D^{(kk)}$ is the Jacobian subdeterminant for all nodes except X_k. Because there are two chains from X_k to X_q, the derivative of p_{qk} is a sum of two terms:

$$\frac{\mathrm{d}p_{qk}}{\mathrm{d}x_k} = w_1P_1 + w_2P_2, \tag{A13}$$

where P_1 and P_2 are the two chain products, and w_1 and w_2 their weights (Plahte et al., 2013). When there is no feedback loop in the system, only the diagonal elements in J stemming from the term $-\gamma_i x_i$ in Equation (A7) contribute to the determinants $D_{V_iV_i}$ and $D^{(kk)}$:

$$\begin{aligned} D_{V_iV_i} &= (-1)^{n-\rho_i}\prod_{j\in V_i}\gamma_j, \\ D^{(kk)} &= (-1)^{n-1}\prod_{j\neq k}\gamma_j. \end{aligned} \tag{A14}$$

Altogether this gives

$$\frac{\mathrm{d}x_q}{\mathrm{d}x_k} = \frac{\mathrm{d}p_{qk}}{\mathrm{d}x_k} = \frac{\gamma_k}{\Gamma_1}P_1 + \frac{\gamma_k}{\Gamma_2}P_2, \tag{A15}$$

where Γ_1 and Γ_2 are the products of the γ_j in the two chains, respectively. The chain products P_1 and P_2 depend on the genotype g_k of X_k as well as on the genotypic background $g^{(k)}$, but their signs S_1 and S_2 are invariant under genotypic variation. It is easy to see that a negative autoregulatory loop, which is a common feature in gene regulatory networks, would not invalidate the conclusion, but a positive autoregulatory loop might.

If the FFL is incoherent, P_1 and P_2 have opposite signs, implying that the sign of $\mathrm{d}x_q/\mathrm{d}x_k$ may vary. If the FFL is coherent, however, no order-breaking can occur.

If X_k is upstream relative to the initial node X_{init} of the FFL, it follows from the above section on networks without loops that there will be no order-breaking in X_{init}, and the above argument is still valid.

MORE GENERAL PHENOTYPES

In real life, relevant phenotypes are not direct gene products, but rather functions of the concentrations of one or several gene products. Let the phenotype $G(g)$ be a function of $x_U(g)$, $G = h(x_U(g))$, where U is a subset of N, and assume that for any $u \in U$, $\partial h/\partial x_u$ has fixed sign for all genotypes. To analyse this case we extend the original system Equation (A2) to

$$\begin{aligned} &\mu_i^{a_i}r_i^{a_i}(x(g)) + \mu_i^{b_i}r_i^{b_i}(x(g)) - x_i(g) = 0, \quad i = 1, \ldots, n, \\ &h(x_U(g)) - x_{n+1} = 0, \end{aligned} \tag{A16}$$

and apply our above results to this system, in which $G(g) = x_{n+1}(g)$, i.e. $q = n + 1$. If there are two nodes among X_U which have a common predecessor X_k, then there will exist two chains from X_k to X_{n+1}. These two chains constitute a feedforward loop with X_{n+1} as final node. If this FFL is incoherent, order breaking due to genetic variation in X_k may occur even if there is no order breaking in the original system comprising the nodes $X_1, \ldots, X_n$. If the FFL is coherent, order breaking only occurs if it occurs in the original system.

REFERENCES

Plahte, E., Gjuvsland, A. B., and Omholt, S. W. (2013). Propagation of genetic variation in gene regulatory networks. *Phys. D* 256–257, 7–20. doi: 10.1016/j.physd.2013.04.002

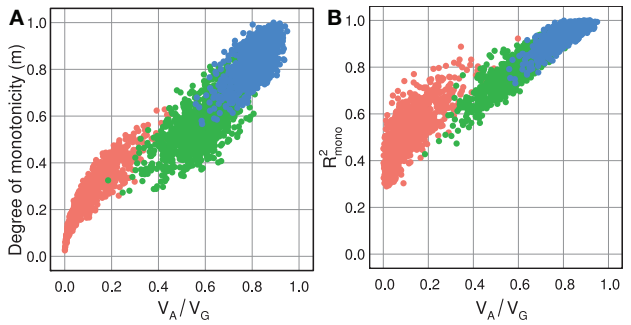


FIGURE A1 | Measures of monotonicity vs. additivity of GP maps with three loci. Scatterplots showing V_A/V_G from unweighted regression vs. **(A)** degree of monotonicity (m) and **(B)** R^2_{mono}. Red dots show 1000 random three-locus GP maps, blue dots show the same 1000 GP maps after sorting to introduce order-preservation for 1 locus while green dots show the same 1000 GP maps after sorting to introduce order-preservation for 2 loci [see Model and Methods in Gjuvsland et al. (2011)].

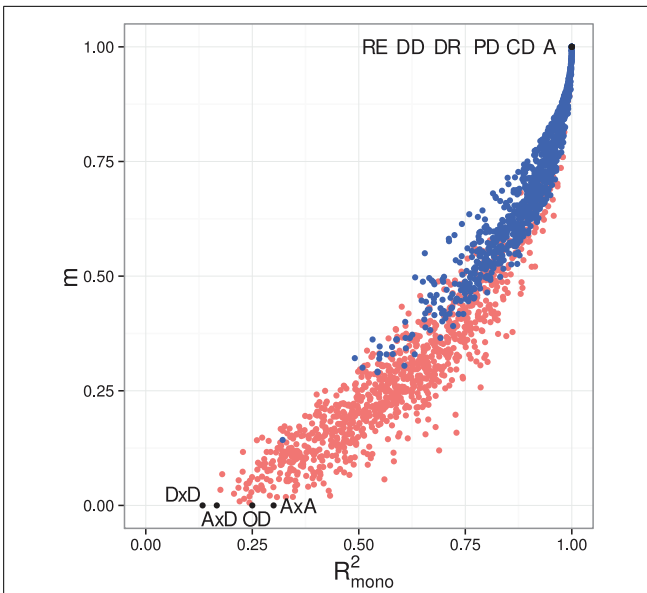


FIGURE A2 | Comparing measures of monotonicity GP maps. Scatterplots showing degree of monotonicity (m) vs. R^2_{mono}. Black dots correspond to the maps shown in **Figure 1**. Red dots show 1000 random two-locus GP maps, while blue dots show the same 1000 GP maps after sorting to introduce order-preservation for 1 locus [see Model and Methods in Gjuvsland et al. (2011)].

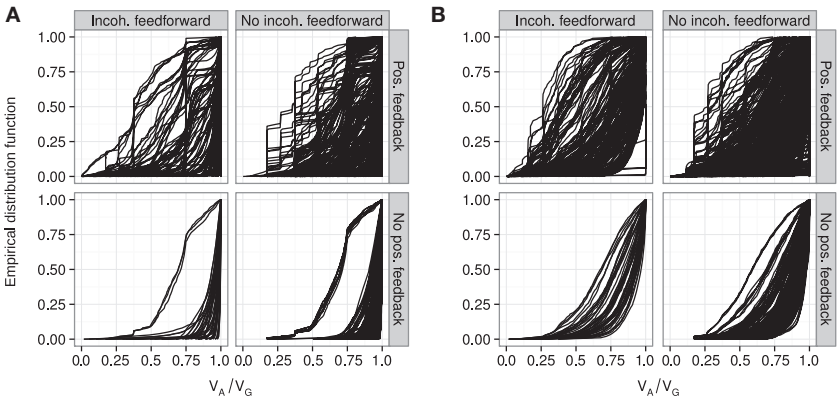


FIGURE A3 | Empirical distribution functions for additivity of GP maps. Summary of V_A/V_G from unweighted regression for all motifs for which at least 100 (out of 1000) Monte Carlo simulations lead to GP maps with non-negligible phenotypic variation (see Models and Methods section "Gene regulatory network simulations," for detailed criteria). Results are shown for 1013 motifs with a genotype-to-parameter map without pleiotropy **(A)** and1090 motifs with a genotype-to-parameter map with pleiotropy **(B)**. Each panel is divided into 4 subplots containing classes of motifs based on the presence/absence of incoherent feedforward and positive feedback loops, see **Table 1** for the number of motifs in each class. Each curve shows, for a single motif, the empirical distribution function value (y-axis) of V_A/V_G from unweighted regression for all GP maps (x-axis).

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 14 July 2014
doi: 10.3389/fgene.2014.00198

Estimating directional epistasis

Arnaud Le Rouzic*

Centre National de la Recherche Scientifique, Laboratoire Évolution, Génomes, et Spéciation, UPR 9034, Gif-sur-Yvette, France

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Michael Kopp, Aix-Marseille University, France
Janna Lynn Fierst, University of Oregon, USA

***Correspondence:**
Arnaud Le Rouzic, Centre National de la Recherche Scientifique, Laboratoire Évolution, Génomes, et Spéciation, Avenue de la Terrasse, Bâtiment 13, 91198 Gif-sur-Yvette, France
e-mail: arnaud.le-rouzic@legs.cnrs-gif.fr

Epistasis, i.e., the fact that gene effects depend on the genetic background, is a direct consequence of the complexity of genetic architectures. Despite this, most of the models used in evolutionary and quantitative genetics pay scant attention to genetic interactions. For instance, the traditional decomposition of genetic effects models epistasis as noise around the evolutionarily-relevant additive effects. Such an approach is only valid if it is assumed that there is no general pattern among interactions—a highly speculative scenario. Systematic interactions generate directional epistasis, which has major evolutionary consequences. In spite of its importance, directional epistasis is rarely measured or reported by quantitative geneticists, not only because its relevance is generally ignored, but also due to the lack of simple, operational, and accessible methods for its estimation. This paper describes conceptual and statistical tools that can be used to estimate directional epistasis from various kinds of data, including QTL mapping results, phenotype measurements in mutants, and artificial selection responses. As an illustration, I measured directional epistasis from a real-life example. I then discuss the interpretation of the estimates, showing how they can be used to draw meaningful biological inferences.

Keywords: epistasis, genetic effects, estimation, statistics, evolution, multilinear model

1. INTRODUCTION

An ability to understand and predict how genes affect morphological, physiological, and behavioral characteristics is of crucial importance in biology. This also poses a considerable challenge, given the complexity of the genetic architecture of quantitative traits (Flint and Mackay, 2009). This complexity is not only due to the large number of genetic, environmental, and physiological factors involved, but also to their multiple and nonlinear interactions. In particular, it was noticed very early in the history of genetics that the same genetic change often produces differing effects depending on the genetic background of the experimental species, population, or individual (Phillips, 1998; Wade et al., 2001; Phillips, 2008). The biological consequences of this phenomenon, known as "epistasis," have triggered a considerable amount of discussion. A whole century of active research in genetics and molecular biology has revealed the ubiquity of epistatic interactions associated with the organization of biological systems as networks of interacting molecules (Omholt et al., 2000). However, we are still far from being able to integrate epistasis into a consensual, explicit, and predictive theoretical framework.

In the classical analysis of genetic variance (Fisher, 1918), epistasis is considered as a source of noise. Most epistatic effects are not transmitted from parent to offspring, and therefore, are not involved in the response to natural or artificial selection. Epistatic variance—the contribution of epistasis to genetic variance in a population—can be calculated (Cockerham, 1954; Kempthorne, 1954; Lynch and Walsh, 1998; Álvarez-Castro and Carlborg, 2007; Gjuvsland et al., 2007), but is almost meaningless in terms of predicting the genetic properties of a population (Barton and Turelli, 2004; Hansen, 2013; Álvarez-Castro and Le Rouzic, 2014), and may be negligible compared to evolutionarily-relevant additive genetic variance (Hill et al., 2008; Hemani et al., 2013).

Another idea, which has become popular only in recent decades, is that epistasis matters because of its capacity to affect additive variance rather than because of its contribution to interaction variance (Cheverud and Routman, 1995). In an epistatic genetic architecture, the effects of alleles on the phenotype depend on the genetic background. Accordingly, changes in the genetic background promoted by genetic drift (Goodnight, 1987, 1988; Barton and Turelli, 2004; Turelli and Barton, 2006; Álvarez-Castro et al., 2009; Jarvis and Cheverud, 2009) or by selection (Carter et al., 2005; Hansen et al., 2006; Hallander and Waldmann, 2007; Le Rouzic et al., 2013) may reveal, hide, or revert allelic effects, and thus significantly affect the genetic variance.

1.1. DIRECTIONAL EPISTASIS

Epistasis can only exert a significant long-term influence on populations if individual epistatic effects do not tend to cancel out each other, i.e., if a general pattern emerges. The most obvious pattern is the directionality of epistasis, the fact that genetic interactions can be biased toward either high or low phenotype values. Estimates of directional epistasis allow to make useful predictions about the evolutionary potential of populations: if additive genetic variance is a measure of evolvability (Houle, 1992; Hansen et al., 2011), then the directionality of epistasis is a measure of genetic architecture asymmetry, i.e., how evolvability is influenced by the direction of evolution. When epistasis is positive, evolution is easier in the direction of high, rather than low, phenotypic values (because additive genetic variance tends to increase with the phenotypic value). In contrast,

negative epistasis favors evolution toward low phenotypic values.

In spite of its predictive and descriptive value, directional epistasis is rarely reported for quantitative characters (Pavlicev et al., 2010). This can be attributed to two main factors: (i) many (if not most) quantitative geneticists are used to measuring epistasis via epistatic genetic variance, in spite of its marginal interest, and (ii) very few statistical or computational tools have been devised for measuring directional epistasis. The aim of this article is to present several methods for estimating directional epistasis from genetic and phenotypic data, and to propose accessible statistical procedures for computing epistasis. Several such methods will be illustrated from a real-life biological example, the genetic architecture of bodyweight in chicken, which displays a clear and consistent signal of positive epistasis. The data is based on a long-term artificial selection experiment on chicken body weight, and features (i) times series of the phenotypic response to selection, (ii) Quantitative Trait Locus (QTL) mapping data from a cross between the divergent lines, and (iii) minimal line-cross information (means of F_1 and F_2 populations) from the QTL setting.

1.2. GENETIC MODELS

In general, measuring the directionality of epistasis requires a model of genetic effects, i.e., a mathematical description of the relationships between the data (for instance, individual genotypes or phenotypes) and parameters to be estimated. The desirable properties for a "good" model of genetic effects depend on both the biological question and the nature of the data, and have resulted in rewarding (and sometimes conflictual) discussions (Cheverud and Routman, 1995; Hansen and Wagner, 2001b; Kao and Zeng, 2002; Yang, 2004; Zeng et al., 2005; Wang and Zeng, 2006; Álvarez-Castro and Carlborg, 2007; Aylor and Zeng, 2008; Hansen, 2014).

Genetic models can be conveniently divided into physiological and statistical models (Cheverud and Routman, 1995). In physiological (or functional: Hansen and Wagner, 2001b) models, genetic effects are described relative to a reference genotype, which can be arbitrary (for instance, one of the parental strains in an intercross) or conventional (typically, the wild genetic background). Functional models are generally rooted in traditional Mendelian genetics, in which a limited number of genotypes are experimentally generated and compared to reference strains. In contrast, statistical models quantify genetic effects in polymorphic populations across multiple genotypes. They are derived from the classical decomposition of genetic variance. Statistical genetic effects depend on allelic frequencies, and thus change when populations evolve; they provide a population-specific description of the genotype-to-phenotype map. In spite of obvious historical and conceptual divergences, it is sometimes possible to express both functional and statistical models in common mathematical frameworks, and to transform functional into statistical estimates (and *vice versa*) by means of "change of reference" operations (Hansen and Wagner, 2001b; Álvarez-Castro and Carlborg, 2007; Le Rouzic and Álvarez-Castro, 2008).

With respect to epistasis, another useful distinction can be made between unidimensional and multidimensional models (Kondrashov and Kondrashov, 2001; de Visser et al., 2011). Unidimensional epistasis describes the general curvature of the genotype-phenotype map, and can be interpreted as the average effect of allelic substitutions that would be observed if all loci were exchangeable. Multidimensional epistasis accounts for the complexity of the genotype-phenotype relationship, by characterizing all pairs of loci that have a specific epistatic effect. While directional epistasis is unidimensional by definition, it can be measured based on either unidimensional or multidimensional models.

Several models of directional epistasis will be reviewed below, starting from the multilinear model of epistasis, originally functional and multidimensional, which has been extended toward statistical and unidimensional formulations. I will then present and discuss alternative functional unidimensional models that are commonly used to measure epistasis for fitness, and show how they can be applied to quantitative characters.

2. MULTILINEAR EPISTASIS

2.1. THE MULTILINEAR MODEL OF GENETIC INTERACTIONS

2.1.1. General framework

The multilinear model of genetic interactions developed by Hansen and Wagner (2001b) extends and makes explicit the concept of directional epistasis in quantitative genetics, and makes it possible to build genotype-to-phenotype maps implementing directional epistasis. In its original multidimensional form, the model expresses the phenotype z as a multilinear function of the genotype G of an individual. For two loci, labeled "1" and "2" respectively,

$$z_G = z_R + y_{1_R} + y_{2_R} + y_{1_R} y_{2_R} \varepsilon_{12}. \quad (1)$$

Genetic effects are measured relative to an arbitrary reference genotype for which $y_1 = y_2 = 0$, associated with a reference phenotype z_R. The effect of substituting the genotype of interest at locus 1 in the reference genotype R is y_{1_R}, and conversely, y_{2_R} is the effect at locus 2. When introducing the genotype of interest at both loci, in the absence of epistasis, the phenotype is expected to change by $y_{1_R} + y_{2_R}$. Any deviation from this expected additive outcome is attributable to epistasis. The originality of the multilinear model is to assume that this deviation is proportional to the product of allelic effects, the proportionality coefficient ε_{12} quantifying the strength and directionality of epistasis between loci 1 and 2.

The multilinearity arises from the fact that any change in the genotype of a locus when keeping the genetic background constant leads to a proportional change in the phenotype. For instance, Equation (1) can be reformulated as $z_G = a + f y_{1_R}$ (with $a = z_R + y_{2_R}$ and $f = 1 + y_{2_R} \varepsilon_{12}$), illustrating that the genotype-phenotype map is always linear with respect to single genotypes (**Figure 1**).

The epistatic coefficient, ε_{12}, is expressed in terms of inversed phenotypic units (e.g., if the trait is measured in cm, ε will be in cm^{-1}), which is not intuitive and does not allow comparisons between traits. Hansen and Wagner (2001b) suggest measuring epistasis by computing epistatic factors, $f_1 = 1 + y_2 \varepsilon_{12}$ and $f_2 = 1 + y_1 \varepsilon_{12}$, which quantify how much locus 1 is affected

by locus 2, and *vice versa*; $f = 1$ implies no epistasis, $f < 1$ negative (antagonistic) epistasis, and $f > 1$ positive (synergistic) epistasis.

2.1.2. Statistical formulation

The multilinear model is built as a functional model, since it defines genetic effects relative to a reference genotype, but a "change of reference" tool can be used to recompute genetic effects in any genotype or weighted combination of genotypes. When genetic effects are calculated relative to the average genotype of a population, the marginal contributions of individual loci coincide with additive effects, and the model can be considered to be statistical.

The multilinear model can also be used as a local approximation on a non-multilinear genotype-phenotype map. There are various ways of generating genotype-phenotype maps, which are multidimensional mathematical functions $g(y_1, y_2, \ldots, y_n)$ that provide a deterministic phenotypic value for a series of genotypic values y_i at n loci. Such mathematical maps are often defined in theoretical work intended to explain the evolution of populations in complex genetic landscapes. Furthermore, even if the lack of large empirical genotype-phenotype data sets means that it is not yet realistic to attempt to do so, it is in principle possible to fit smooth surfaces (such as multidimensional splines) to experimental measurements, and thus generate models of genetic landscapes that could be analyzed mathematically (and tested empirically).

In any case, the multidimensional directional epistasis coefficients ε_{ij}, which measures the curvature of the genotype-phenotype function between loci i and j, can be directly quantified as $\varepsilon_{ij} = D^2_{ij}/D_iD_j$, where $D_i = \partial g/\partial y_i$ is the value of the first partial derivative of function g taken at the reference point, and $D^2_{ij} = \partial^2 g/\partial y_i \partial y_j$ is the mixed partial derivative (the curvature of the function g across both loci). This result illustrates the fact that the multilinear model is similar to a Taylor expansion of the genotype-phenotype map that ignores intra-locus curvature (Hansen and Wagner, 2001b) (see Appendix I and **Figure 2**).

2.1.3. Composite directional epistasis

The original multilinear model is multidimensional, as it involves as many ε_{ij} parameters as pairs of loci. A unidimensional (and

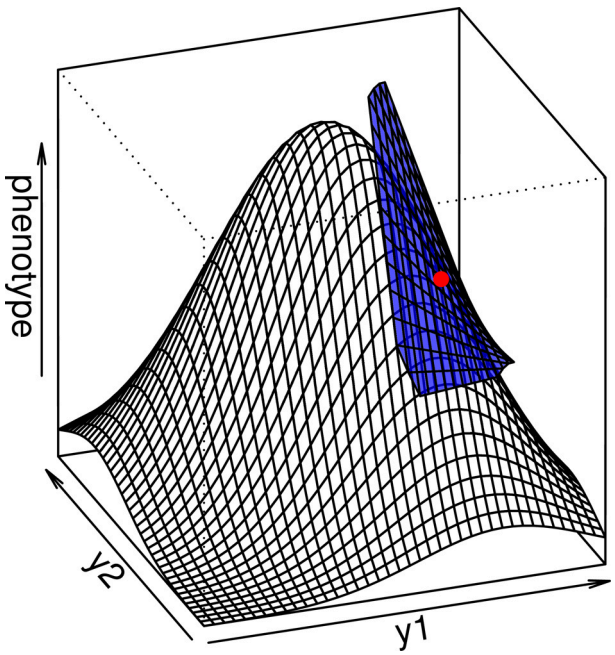


FIGURE 2 | The multilinear model (blue surface) is a local approximation of the interlocus curvature in a complex genotype-phenotype map. When the average genotype is chosen as the reference (red point), the multilinear approximation is able to predict the evolutionary properties of the population in a more precise way than the additive model.

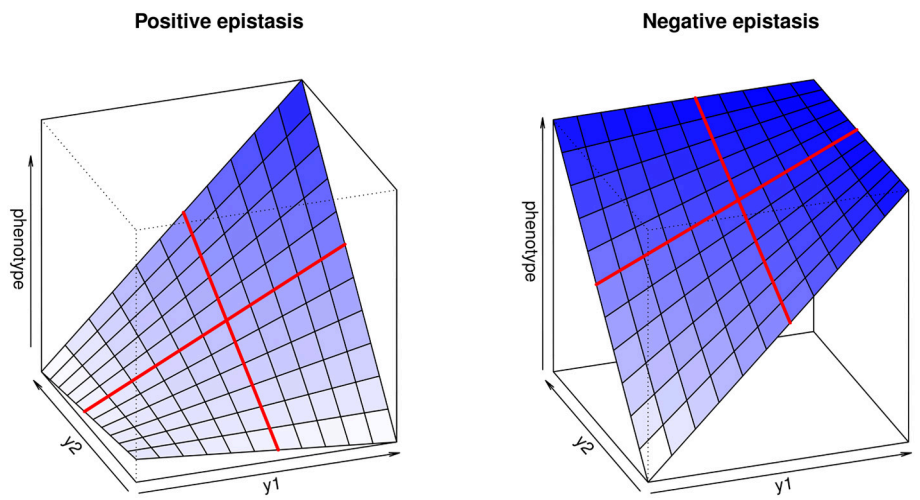


FIGURE 1 | Multilinear genotype-phenotype maps for two loci, illustrating positive (synergistic) and negative (antagonistic) epistasis. y_1 and y_2 represent the genotype values at both loci. The red lines highlight the multilinearity of the model: if the genetic background is kept constant, phenotype change depends linearly on the genotype at each locus.

statistical) version of the model was proposed in Carter et al. (2005), with the composite directional epistasis coefficient ε_c calculated as the average ε_{ij} coefficient weighted by the additive genetic variance explained by each pair of loci:

$$\varepsilon_c = \frac{\sum_i \sum_{j \neq i} V_{A_i} V_{A_j} \varepsilon_{ij}}{\sum_i \sum_{j \neq i} V_{A_i} V_{A_j}}. \quad (2)$$

Both uni- and multi-dimensional versions of the model can be extended to higher orders of interactions and to multiple traits (Hansen and Wagner, 2001b).

2.2. DIRECTIONAL EPISTASIS FROM PHENOTYPIC DATA

2.2.1. Response to artificial selection

Directional epistasis affects evolution, as it changes the amount of genetic variation available depending on the direction of phenotypic change (Hansen et al., 2006). For instance, selection in the direction of positive epistasis tends to increase the frequency of synergistic genetic interactions, thus enhancing the effect of selection. In contrast, selection in an antagonistic system decreases the genetic variance, and thus decreases the selection response. These effects can be experimentally observed, especially with bidirectional artificial selection responses, since they are expected to generate asymmetric responses in up- and down-selected lines.

2.2.1.1. Theoretical framework. It is possible to model the expected impact of directional epistasis on genetic variance and to predict the difference between up- and down-selected lines as a function of the epistatic coefficients. Using a series of simplifying assumptions detailed in Appendix II, the selection response under a constant selection gradient after t generations is expected to be:

$$\begin{aligned} \mu_t &\simeq \mu_0 - \frac{\log(1 - 2\Delta_{\mu_0}\varepsilon t)}{2\varepsilon} \\ &\approx \mu_0 + \Delta_{\mu_0} t + \varepsilon \Delta_{\mu_0}^2 t^2 + \ldots, \end{aligned} \quad (3)$$

where μ_0 is the initial mean phenotype, Δ_{μ_0} is the initial selection response (after the first generation), and ε is the directionality of epistasis. The second part of the equation is the second-order Taylor approximation around $t = 0$, illustrating the linear selection response expected by the traditional breeder's equation ($\Delta_{\mu_0} t$), and how directional epistasis appears as a quadratic term. Here, ε is the unidimensional directional epistasis, and thus corresponds to ε_c in Equation (2).

A convenient way to estimate directional epistasis from bidirectional selection responses is to compute the up/down asymmetry through the average selection response, $A(t) = \frac{1}{2}(\text{up}(t) + \text{down}(t))$ (**Figure 3**). If epistasis is directional and relatively weak ($\Delta_{\mu_0}\varepsilon \ll 1$), $A(t)$ changes approximately with t^2, such that $A(t) \simeq \varepsilon \Delta_{\mu_0}^2 t^2$. It is thus possible to estimate Δ_{μ_0} as the slope at origin of the selection response, and then ε through a quadratic regression on the average up/down response. Including the effects of e.g., inbreeding, linkage disequilibrium, or canalization, is possible, but requires to numerically maximize the likelihood of complex models. This can be done with the software package `sra` for `R`, described in Le Rouzic et al. (2011).

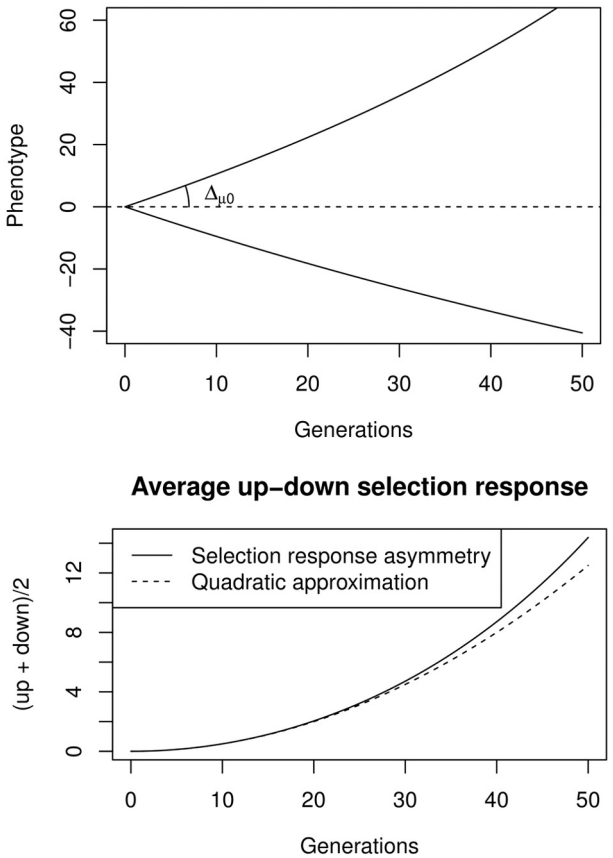


FIGURE 3 | Top: Theoretical response to bidirectional constant selection under positive directional epistasis ($\varepsilon = +0.005$, $\Delta_{\mu_0} = 1$). **Bottom:** the selection response is asymmetric, and the up-down average increases almost quadratically with time, the quadratic coefficient being $\varepsilon \Delta_{\mu_0}^2$.

2.2.1.2. Example: artificial selection on body weight. For more than 50 years, two chicken (*Gallus gallus*) lines were selected for high and low body weight at 56 days, respectively (Siegel, 1962; Liu et al., 1994; Dunnington and Siegel, 1996). The experiment is still ongoing; here, I consider the latest phenotypic results available (54 generations, Dunnington et al., 2013). For simplicity, only the time series of mean phenotypes are considered, although some variance estimates were also available in this case.

The impact of artificial selection was considerable (**Figure 4**). In the high-selection line, the body weight at 8 weeks rose from 800 g (male-female average) to 1650 g. In the low-selected line, the average body weight decreased to around 150 g, leading to an impressive order-of-magnitude difference between high- and low-selected lines, well beyond the differences usually observed between closely-related species, and spanning more than one third of the relative weight diversity in the entire 20 Myr-old Galliformes order. The selection response was asymmetric: although the selection strength was identical in both lines, progress was slower in the low line. This can easily be attributed

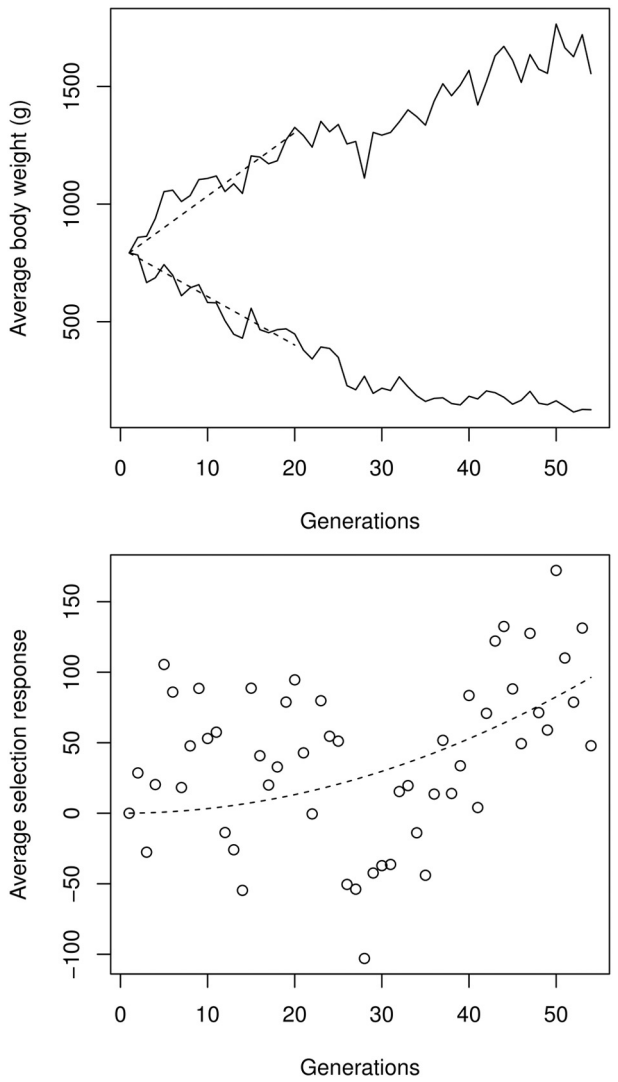


FIGURE 4 | Top: male-female average experimental selection response on chicken bodyweight, digitalized from Figures 1, 2 in Dunnington et al. (2013). The initial selection response Δ_{μ_0}, estimated by a linear regression over the first 20 generations (dashed segments), was 25.6 g per generation in the high line, and −19.6 g per generation in the low line. **Bottom:** quadratic regression on the up- and down-selection average, illustrating the cumulative effect of directional epistasis. The quadratic coefficient (which is an approximation of $\Delta^2_{\mu_0}\varepsilon$), estimated by a non-linear, least-square regression, was 0.033 g per generation squared.

to epistasis, given the expected differences in the genetic backgrounds of 1500 vs. 150 g birds.

Using the procedure described in Equation (3), the strength of directional epistasis could be estimated from a quadratic regression over the high-low asymmetry. Estimating the initial selection response at around $|\Delta_{\mu_0}| = 22.6$ g per generation on average, directional epistasis is $\varepsilon \simeq +6.6 \times 10^{-5}\,\mathrm{g}^{-1}$. Although apparently small, this figure is statistically significant and generates cumulative effects on genetic architectures: Any phenotypic change corresponding to the initial (first-generation) selection response induces an increase of allelic effects of 0.15% in the high line, and decreased accordingly in the low line. The same allele is thus expected to display a >10% difference in the two extreme genetic backgrounds, representing weak, but non-negligible, epistasis.

Of course, this estimate relies on major assumptions about the underlying process. Several genetic or non-genetic factors other than epistasis could affect the available genetic variance, and thus bias ε. For instance, the quadratic approximation relies on the hypothesis that the selection gradient is constant over the entire time series, whereas in fact we know from e.g., Dunnington et al. (2013) that the selection intensity actually increases with time. Meanwhile, the reduced population size in the experiment necessarily generated a significant amount of inbreeding (even with a carefully-designed breeding scheme), which decreases the variance due to genetic drift. However, these mechanisms are unlikely to generate misleading estimates of ε, since (i) they affect both the up and down lines in the same way, and so cannot generate any asymmetry, and (ii) they tend to offset each other, as the selection strength increases while the genetic variance decreases.

More worrisome is the possibility of uncontrolled natural selection in the low line. A fraction of the smallest birds appeared to be sterile or unviable, which could contribute to the slowing-down of the response. Such a mechanism could generate an asymmetric response, and thus spurious positive estimates of the epistatic coefficient. Nevertheless, this seems rather unlikely, given the behavior of the twelve relaxed selection lines presented in Dunnington et al. (2013). Indeed, when selection was stopped in both lines, the populations did not tend to evolve back to the original phenotype, as would have been expected if natural selection was preventing the population from responding to artificial selection. The phenotypic data therefore seems to be compatible with a genetically-driven asymmetry, due to smaller allelic effects in low-weight chickens (i.e., positive epistasis).

2.2.2. Line-cross analysis

With the improvement in sequencing and genotyping technologies, the phenotype-based methods developed and used by quantitative geneticists for most of the 20th century to investigate genetic architectures without resorting to genotype data are currently losing popularity. However, they are still both elegant and informative, especially when used to estimate general properies of populations such as unidimensional directional epistasis. One of the most powerful (and simple) of these biometric methods consists of crossing individuals or strains of interest in order to generate hybrid and backcross populations, from which the phenotypic means and variances can be determined. The knowledge of the transmission mechanisms of genetic factors from parents to offspring makes it possible to disentangle the impact of additive, dominance, and epistatic effects on the genetic differences between the original individuals (Lynch and Walsh, 1998 p. 205).

A set of equations that can be used to compute additive, dominance, and directional epistatic effects from parental, intercross, and backcross populations are provided in Hansen and Wagner (2001b) (see Demuth and Wade, 2005, for an alternative model). Directional epistasis is unidimensional, and thus corresponds to the ε_c parameter of Equation (2). Below, a slightly different parameterization will be used, in which both parental populations

are separated by four additive effects, so that the model is identical to a 2-locus QTL effect model in a diploid species. The model was set up so that genetic effects cancel out in the F_2 population, but a different reference point can be chosen (using the genetic effect matrices provided in, e.g., Álvarez-Castro and Carlborg, 2007). Average phenotypes for both parental populations (P_1 and P_2) and the first two intercross populations F_1 and F_2 can be express as functions of four parameters: a reference μ (arbitrarily, the mean F_2), additive and dominance effects A and D, and the directional epistasis coefficient ε.

$$\begin{aligned} P_1 &= \mu - 2A - D + \varepsilon(A^2 + AD + \frac{1}{4}D^2) \\ P_2 &= \mu + 2A - D + \varepsilon(A^2 - AD + \frac{1}{4}D^2) \\ F_1 &= \mu + D + \frac{1}{4}\varepsilon D^2 \\ F_2 &= \mu. \end{aligned} \tag{4}$$

This simple model can be illustrated by the data from the experimental cross between the two chicken strains (Dunnington and Siegel, 1996; Marquez et al., 2010). In this experiment, the two generations of crossing necessary to generate a polymorphic F_2 population for QTL mapping makes it possible to sketch a minimal line-cross analysis. Both parental populations as well as F_1 and F_2 individuals were raised in the same location, with the same food, and at the same density; their average weights at 8 weeks were 170 and 1412 for both parental chicken populations respectively, 650 g for the F_1, and 624 g for the F_2. Both F_1 and F_2 are below the parental arithmetic average (791 g), suggesting the presence of dominance and/or epistatic effects (Álvarez-Castro et al., 2012).

Although not perfect, this setting makes it possible to estimate up to four genetic parameters. Two models, with and without dominance, were tested, and gave very similar results (Equation 4 and **Table 1**). The dominance effect, when estimated, was an order of magnitude below the additive contribution. Epistasis was positive, and of similar magnitude in both models.

2.3. DIRECTIONAL EPISTASIS FROM QTL DATA

Nowadays, data sets often consist of individuals in which both the phenotype and the genotype at loci of interest are known. This is for instance the case after the mapping of Quantitative Trait Loci (QTLs), either by linkage or association methods. Such data sets represent a valuable source of information about epistasis, and in particular about multidimensional epistasis, which can hardly be estimated from phenotypic data.

Table 1 | Epistatic line-cross analysis of the chicken lines.

Effect	No dominance	Dominance
Reference μ	637 g	624 g
Additive A	310 g	318 g
Dominance D	-	26 g
Directional epistasis ε	1.6×10^{-3} g^{-1}	1.9×10^{-3} g^{-1}

The full model (involving dominance) has no degree of freedom, so that statistical errors cannot be estimated.

2.3.1. Linear and multilinear models of genetic effects

In most cases, QTL mapping procedures only focus on marginal (additive and dominance) effects, and do not explicitly consider genetic interactions (Carlborg and Haley, 2004). However, epistasis may be of major interest, both for improving QTL detection (Carlborg et al., 2003, 2004, 2006), and for the biological interpretation of the genotype-phenotype relationship (Malmberg and Mauricio, 2005; Le Rouzic et al., 2007, 2008). Mapping procedures accounting for epistasis generally rely on components of the interaction variance (Cockerham, 1954; Kempthorne, 1954; Lynch and Walsh, 1998), which makes it necessary to estimate four genetic effects for each pair of loci (additive-by-additive, additive-by-dominant, dominant-by-additive, and dominant-by-dominant statistical effects). More recently, "variance QTL" approaches have been proposed to map loci involved in various kinds of interactions, including gene-gene and gene-environment interactions (Rönnegård and Valdar, 2012). Until recently, there was no QTL mapping method based on directional epistasis (Slatkin and Kirkpatrick, 2012), and estimation from genotype-phenotype data usually relied on model fitting on a predefined set of candidate loci (Cheverud et al., 2001; Le Rouzic et al., 2008; Shao et al., 2008; Pavlicev et al., 2010; Jarvis and Cheverud, 2011).

The traditional genetic regression model, ignoring dominance (and dominance-related epistatic components), can be written as:

$$P_{y_1,y_2} = \mu + \alpha_1 S_1 + \alpha_2 S_2 + \alpha\alpha_{12} S_{12}. \tag{5}$$

This model has 4 parameters for a pair of loci: μ is the intercept of the model (reference point), α_1 and α_2 are the additive effects for both loci, and $\alpha\alpha_{12}$ — a traditional (and probably unfortunate) notation, not to be confused with the product $\alpha \times \alpha_{12}$ — is the additive-by-additive effect. The S coefficients determine the genetic model, i.e., the weights of the genetic effects for each genotype. For instance, consider a haploid two-locus two-allele system with the reference genotype (arbitrarily) set to A_1B_1. In the reference genotype, all S coefficients are set to 0 (μ, the reference point, thus corresponds to the intercept of the model). For genotype A_1B_2, $S_1 = 0$, $S_2 = 1$ (because 1 effect α_2 has been added to the model, given the substitution of a B_2 allele), and $S_{12} = 0$. In genotype A_2B_2, $S_1 = 1$, $S_2 = 1$, and $S_{12} = 1$, reflecting the possibility of an interaction between A_2 and B_2 alleles. Of course, different reference points can be chosen, including mixtures of genotypes in specific frequencies (such as in the F_2 model, considering even allelic frequencies and Hardy-Weinberg proportions). The models becomes more complex with diploid genotypes (which include dominance effects), but the principle remains the same. Below, I used the model "NOIA" proposed by Álvarez-Castro and Carlborg (2007), which has some interesting statistical features. In particular, the model is orthogonal (provided there is no linkage disequilibrium) even if the population is not at Hardy-Weinberg proportions. In "NOIA," the S coefficients are stored as a genetic design matrix, and the model can be extended

(to include more alleles and/or more loci) using simple matrix algebra.

It is possible to modify the above framework to estimate directional epistasis. The strategy proposed by Le Rouzic and Álvarez-Castro (2008) is based on a non-linear, least-square regression, very similar to the framework proposed in Equation (4) for the analysis of line crosses: the model explicitly decomposes the epistatic parameter as a multilinear combination of additive effects, assuming that $\alpha\alpha_{ij} = \alpha_i \times \alpha_j \times \varepsilon_{ij}$:

$$P_{y_1,y_2} = \mu + \alpha_1 S_1 + \alpha_2 S_2 + \alpha_1\alpha_2\varepsilon_{12} S_{12}. \quad (6)$$

This setting can easily be extended to account for dominance and higher-order epistasis (Álvarez-Castro and Carlborg, 2007; Le Rouzic and Álvarez-Castro, 2008; Pavlicev et al., 2010). When ε_{ij} is estimated for each pair of loci, the model describes multidimensional epistasis. There are two distinct ways to estimate unidirectional epistasis from this setting. The first method is to assume that ε is identical between loci, i.e., replacing ε_{ij} by a constant ε in Equation (6). The second strategy is to estimate independent ε_{ij} values for each pair of loci, and to compute the composite epistasis ε_c using Equation (2). This last strategy is more theoretically-grounded than the former, but it rapidly becomes impractical when the number of loci increases: the number of interactions increases quadratically with the number of loci, which reduces the precision of pairwise interaction estimates.

2.3.2. Application to QTLs for body weight

Individuals from both the high and low chicken lines were intercrossed at generation 46, to form the F_1 and F_2 populations described above. The 795 surviving individuals from the F_2 population were phenotyped for various characters and genotyped for 145 genetic markers on 25 chromosomes. The QTL mapping analysis identified 6 significant loci (four major loci and two of lesser effect). These significant loci combined explained around 10% of the phenotypic variance, and strong epistatic interactions have been reported among them (Carlborg et al., 2006; Le Rouzic et al., 2007; Álvarez-Castro et al., 2012). For the sake of both simplicity and statistical power, only the four major QTLs are considered in the subsequent analyses.

There are 24 second-order epistatic interactions between four loci (6 additive-by-additive, 6 dominance-by-dominance, and 12 additive-by-dominance interactions). It is possible to estimate all of them using a model performing the traditional decomposition of genetic effects (here, I used the software package `noia` for `R`, Le Rouzic and Álvarez-Castro, 2008), but interpreting these 24 independent epistatic estimates is complicated: in spite of the large sample size (around 800 individuals), only 4 (out of 24) epistatic estimates reached the 5% p-value threshold, and none remained statistically significant after correction for multiple-testing. There were no obvious signs of directional epistasis (11 positive estimates out of 24), even when focusing on additive-by-additive epistasis (3 positive estimates out of 6).

Fitting a unidimensional multilinear model of epistasis leads to a much more conclusive analysis. The estimated constant ε coefficient is positive ($\varepsilon = +0.057\,\mathrm{g}^{-1}$). The weighted composite parameter, calculated from Equation (2), is also positive and of the same order of magnitude ($\varepsilon_c = +0.020\,\mathrm{g}^{-1}$). The multilinear model fits better than the traditional genetic-effects model with pairwise epistasis, outperforming it by 13.5 AIC units (ΔAIC scores >10 can be considered to be conclusive, Burnham and Anderson, 2002). The multilinear model is also considerably better than models without epistasis ($\Delta\mathrm{AIC} = 18.5$). The undisputable statistical superiority of the multilinear model translates into a substantial gain in explanatory power: the four-locus model without epistasis explains only 5.4% of the total phenotypic variance, while the multilinear model explains 7.8%.

3. REGRESSIONS AGAINST THE NUMBER OF MUTATIONS

While it is particularly rare to find estimates of directional epistasis for quantitative characters in general (Pavlicev et al., 2010), the sign of epistasis has been frequently estimated for fitness. The importance of directional epistasis for the logarithm of fitness has now been fully acknowledged by evolutionary biologists, as it affects the evolution of sex, recombination, mutation rates, and other related phenomena (Phillips et al., 2000). Here I will review two models frequently used in this context, and show how they can be modified to fit other quantitative traits. According to the previous definitions, these models are both functional and unidimensional, as they estimate directional epistasis with reference to the "wild type" with no mutations.

3.1. MODEL DESCRIPTION

A common way to estimate directional epistasis for (log) fitness is a "power" (or "multiplicative") model $W = \alpha n^{\beta}$ (illustrated in **Figure 6**), where W stands for the log-fitness, α is the effect of a single mutation, n is the number of mutations, and β measures directional epistasis. The model is based on the fact that the fitness of the reference individual or strain ($n = 0$) is 1, so that the intercept of the model is $\log(1) = 0$ by construction. Fitness in single mutants ($n = 1$) is not affected by epistasis, which makes it possible to estimate α. Epistasis appears for $n \geq 2$, generating deviations from linearity. $\beta > 1$ represents positive epistasis, while $\beta < 1$ stands for negative epistasis. The parameters of the model are usually estimated through non-linear regressions (least squares) or by non-linear generalized model approaches (maximum likelihood).

An alternative setting is the quadratic model $W = -(\alpha n + \frac{1}{2}\beta' n^2)$ (Elena and Lenski, 1997; Kouyos et al., 2007) (for consistency with the literature, I have retained the same notation, although it should be noted that β and β' have different units, and $\beta' > 0$ means positive epistasis). This latter model has some interesting theoretical properties associated with the Gaussian fitness function, and is more firmly grounded in classical population genetics theory (Charlesworth, 1990; Otto, 2007).

Alternative parameterizations of the above models appear in the literature (e.g., estimating $-\alpha$ instead of α, or $\beta - 1$ instead of β, which provides a more straightforward interpretation of "positive" and "negative" epistasis). This framework is generally used in two different experimental contexts: estimating the directionality of deleterious mutations (in which case, $\alpha < 0$, and negative epistasis means that the deleterious mutations act synergistically to decrease fitness), or estimating epistasis among the beneficial mutations accumulated during an artificial evolution

experiment ($\alpha > 0$, and negative epistasis represents the antagonistic effects of mutations) (Lenski et al., 1999; Wilke and Adami, 2001; Maisnier-Patin et al., 2005). These symetric interpretations are arguably confusing, and the literature is not always consistent with regard to the association between the sign of directional epistasis and the synergistic or antagonistic properties of mutations (e.g., Szathmáry, 1993).

3.2. MODEL FITTING

These models are clearly not suited for fitting traditional quantitative genetics data, in which there are no "wild type" or "mutants." However, it is still possible to define the following continuous function for a phenotype P, which behaves in a similar fashion as the power model:

$$P(m) = \begin{cases} \mu + \alpha m^{\beta}, & \text{if } m > 0 \\ \mu, & \text{if } m = 0 \\ \mu - \alpha |m|^{1/\beta}, & \text{if } m < 0, \end{cases} \quad (7)$$

where m is a real number analogous to the "number of mutations" compared to the reference genotype, α and β have the same meaning as in the power model (α is the average effect of the first mutation, and β is the epistatic coefficient, with $\beta = 1$ standing for no epistasis). μ is the intercept of the model, i.e., the phenotype of the "reference genotype." This function is not differentiable at $m = 0$, but this is unlikely to affect the estimates. In order to obtain a proper analogy with traditional quantitative genetics, the mean F_2 (same number of alleles from both parental lines) was chosen as the reference. m, the "number of mutations" parameter, thus stands for the number of additional "high-line" (H) alleles in a genotype compared to the reference. Considering the 4 significant QTLs, $m = 0$ for the reference (mean F_2) genotype (which has 4 low-line alleles and 4 high-line alleles), $m = -4$ in the full low-line genotype (8 alleles from the low-line), and $m = +4$ in the full high-line genotype. An equivalent formulation ($P(m) = \mu + \alpha m + \frac{1}{2}\beta' m^2$) can also be defined for the quadratic model.

Fitting the "continuous power model" of Equation (7) to the data by a non-linear, least-square procedure leads to the following estimates (estimate ± std. err.): $\alpha = 13.0 \pm 5.8$ g; $\beta = 2.18 \pm 0.41$ (**Figure 5**). This is indicative of strong (and statistically significant) positive epistasis. The first allelic substitution in the reference background (average F_2 individual) is thus expected to have an effect of 13 g, the second substitution will affect the phenotype by 45.9 g (two "high" substitutions) or 4.9 g (two "low" substitutions). The epistatic effect is extreme for the fourth substitution, which is predicted to have an effect of 124 g in the "high" direction (i.e., 10 times the estimated effect in the average genetic background) but only 3 g in the "low" direction. The estimate of directional epistasis in the power model is heavily influenced by the few "extreme" genotypes: the 7 individuals with eight "H" alleles are all far above the average, which contributes to the excessive curvature of the genotype-phenotype relationship (**Figure 5**). Yet, epistasis is still present when all extreme genotypes (full homozygotes LL and HH) are removed, with an estimate of $\beta = 1.83 \pm 0.50$.

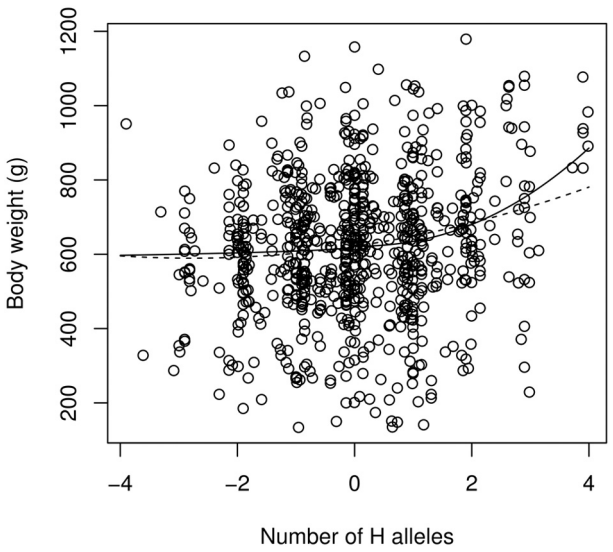


FIGURE 5 | The continuous version of the power model (Equation 7, solid line) and the quadratic model of epistasis (dashed line) applied to the chicken QTL data. The reference genotype contains as many "low" (L) alleles as "high" (H) alleles. The x-axis scales from −4 (LL genotype at all loci) to +4 (HH genotype at all loci). Intermediate numbers of mutations are due to genotype uncertainties when QTLs are not in total linkage disequilibrium with markers.

Estimates from the quadratic model are $\alpha = 23.1 \pm 4.7$ g, and $\beta' = 8.3 \pm 4.0$ g. In spite of the similar notation, β' is not on the same scale as β, and directional epistasis, although significantly positive, is smaller here (the two first allelic substitutions in the direction of higher phenotypes have an effect of 27.3 and 35.6 g respectively, vs. 19.0 g and 10.7 g for one and two substitutions toward lower phenotypes).

4. DISCUSSION

4.1. MODEL COMPARISONS

Although they all provide an estimate of unidimensional directional epistasis, the models reviewed in this paper have been designed to address different questions, and based in different sub-fields of population and quantitative genetics.

The multilinear model provides an explicit description of epistasis between a set of loci, as in classical quantitative genetics models, and can be extended to fit to phenotypic data. On the opposite, both "regression" models suppose that epistatic patterns follow a general function. This incompatibility between models of directional epistasis for fitness and traditional quantitative genetics models is probably an important factor in the lack of experimental measurements of directional epistasis for quantitative traits (Hansen and Wagner, 2001a; Pavlicev et al., 2010).

In addition to the fact that models are not designed to be applied to the same kind of data (the need to compare genotypes to an arbitrary wild type or the assumption of constant mutational effect size are difficult to overcome for quantitative genetics data), models also carry conceptual differences about the nature of epistatic interactions. For instance, the power model

necessarily involves highly complex epistatic interactions (Hansen and Wagner, 2001a). Quantitative genetics rely on linear models of genetic effects, in which interactions are calculated iteratively as the deviation between mutant phenotypes and the sum of lower effect interactions. The multilinear model follows this tradition, and is built as a sum of effects involving one locus (marginal effects), two loci (pairwise interaction effects), three loci, etc. For instance, second-order epistasis is the difference between the double mutant and twice the single mutant effect (**Figure 6**). In contrast, in the power model, there are as many interaction effects as there are mutations, which leads to very complex epistasis. For most realistic values of β ($0 < \beta < 2$), the second- and third-order interactions have opposite effects—in other words, if combining two mutations has antagonistic effects, combining three of them will have synergistic effects (the triple mutant is closer to additivity than predicted by the sum of second-order interactions). Moreover, the magnitude of high-order epistatic effects can represent a substantial fraction of lower-order effects (**Figure 6**), suggesting that combined mutant phenotypes are heavily impacted by the emergent properties of specific combinations of allelic substitutions, and thus difficult to predict from experimental results.

This issue is avoided with the quadratic model, which is limited to interactions between pairs of loci. However, this quadratic model implies that mutational effects can switch signs depending on the genetic background (sign epistasis). This property, which is sometimes perceived as undesirable when considering epistasis for fitness (Wilke and Adami, 2001), could explain the persistence of alternative models. Another side effect of most unidimensional models of epistasis for fitness is that mutations are assumed to be of constant size. Relaxing this assumption significantly alters the evolutionary properties of the system (Butcher, 1995; Otto and Feldman, 1997), casting doubts on the operational meaning of β (or β') parameters.

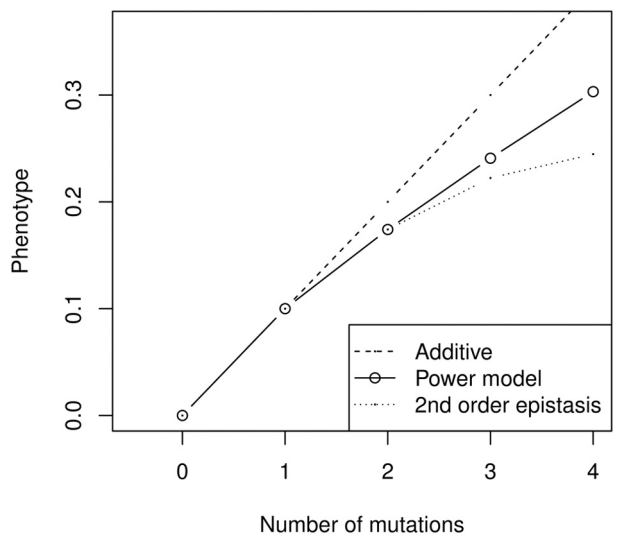


FIGURE 6 | Illustration of high-order epistatic effects in the power model (here with negative epistasis, αn^{β} with $\alpha = 0.1$ and $\beta = 0.8$). The second-order epistatic effect is negative (the power model is always below the additive prediction), but the third-order effect is positive (the power model is always above the quadratic model). The sign of the interactions thus alternates when $\beta < 2$, and their relative size does not decrease rapidly. As a result, the effect of combining several mutants cannot be properly inferred from simpler combinations—for instance, the prediction for four mutants is not much better for the second-order epistatic model than for the additive model, and can even be worse with more substitutions.

4.2. FULL-GENOME EPISTASIS

For most of the 20th century, the concept of genotype-to-phenotype map was mostly virtual, and mainly used for theoretical purposes. The possibility to access complete individual genomes for a reasonable price has not really been anticipated by quantitative geneticists, and we are now in the uncomfortable situation of not being able to properly translate the massive amount of data collected experimentally into ground-breaking theoretical insights. Indeed, it is widely acknowledged that the revolutionary improvement in the quality and quantity of genotypic information has not generated a proportional improvement in our ability to describe the genetic architecture of quantitative traits from genome-wide association studies. This "missing heritability" problem might be partly due to our inability to detect properly epistatic interactions (Maher, 2008; Zuk et al., 2012; Hemani et al., 2013).

Identifying interacting pairs of loci from a genotype-phenotype dataset schematically follows two strategies: (i) combine epistatic and marginal effects while mapping loci, with the hope to increase the genetic signal (Carlborg and Haley, 2004), or (ii) first map loci based on their marginal effects, and estimate epistasis *a posteriori* between pairs of significant loci. Although theoretically elegant, the first strategy generally collapses with high-quality sequencing data because there are so many pairwise combinations to be tested that statistical noise overcomes the genetic signal by orders of magnitude. So far, the second strategy is thus unavoidable for estimating epistasis from high-throughput sequencing data. On the one hand, some epistatic loci will not be detected (in particular, those involved in sign epistasis, which may have no marginal effect). On the other hand, we know from Equation (2) that the impact of loci on the composite epistatic coefficient is weighted by their (marginal) genetic variance, meaning that the loci with no additive effects will not affect directional epistasis. Consequently, estimating epistatic noise in general remains a complex task, and may require further statistical development. When it comes to directional epistasis, focusing on major loci is much less problematic and ensures a proper estimation of this biologically meaningful parameter.

4.3. CONSISTENCY ACROSS ESTIMATES

This paper illustrates the estimation of epistasis directionality by several methods, using independent data describing the same biological system. The various estimates are reported in **Table 2**. The units and the meaning of the epistatic coefficients differ according to the method. In order to facilitate the comparison, an epistatic factor f_{100} is provided. This factor corresponds to the coefficient by which genetic effects change when body weight increases by (arbitrarily) 100 g.

Table 2 | Summary of the directional epistasis estimates from different sources of data and different methods.

Source of data	Method	Estimate	f_{100}
Selection response	Quadratic approximation (Equation 3)	$\varepsilon = 6.6 \times 10^{-5}\,g^{-1}$	1.007
Line cross	Line cross analysis (Equation 4)	$\varepsilon = 1.9 \times 10^{-3}\,g^{-1}$	1.19
QTL	Multilinear regression	$\varepsilon = 5.7 \times 10^{-2}\,g^{-1}$	6.7
QTL	Power model (Equation 7)	$\beta = 2.18$	6.6
QTL	Quadratic model	$\beta' = 8.3\,g$	2.0

Estimates can be compared with the f_{100} factor.

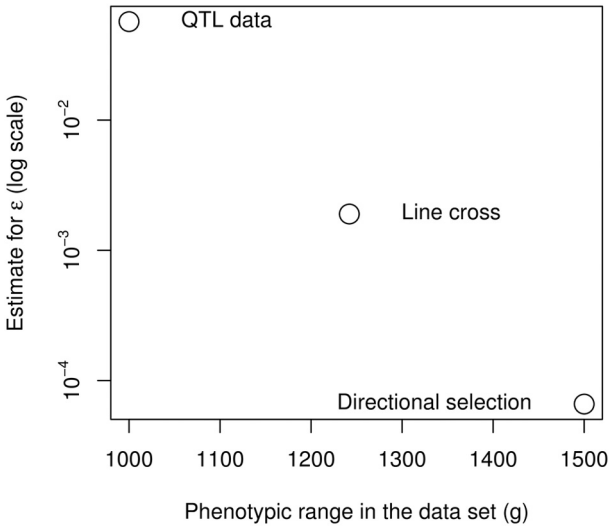


FIGURE 7 | Negative relationship between the span of the phenotypes in the data set and the directional epistasis coefficient.

Directional epistasis estimates are consistently positive, and in most cases statistically significant. This provides strong confirmation that the genetic architecture of the weight differences between the high and low chicken lines is characterized by positive epistasis. However, the epistatic coefficients vary by several orders of magnitude in the different experiments; two categories of estimates can be defined: epistasis is strong when measured from the genotype data (increasing the phenotype by 100 g multiplies the allelic effects by 2 to almost 7), but weaker when measured from phenotype data (increasing the phenotype by 100 g increases allelic effects by 0.7 to 19%).

These measures are not necessarily contradictory, because epistasis can be restricted to a specific subset of the genetic architecture. As the epistatic coefficient measures the "average" curvature of the genotype-phenotype map, it is strongly affected by the nature of the data (and more specifically, the span of the data in terms of number of loci and phenotype range), as it seems to be the case for the chicken bodyweight (**Figure 7**). The extreme epistatic factors measured from the QTL data can be attributed to several factors. The four large-effect QTLs are not a random sample of loci, their effect is statistically inflated by detection bias (the Beavis effect: Beavis, 1994; Xu, 2003), and their strong epistatic interactions remain atypical (Carlborg et al., 2006). Their interaction pattern involves sign epistasis (Le Rouzic et al., 2007), so that additive effects vanish in some genetic backgrounds: increasing a small effect by a large factor does not necessarily mean that the absolute interaction effect is huge. In any case, even if positive epistasis is very strong for the 4 major loci, these QTLs only explain 7% of the total phenotypic variance, and the F_2 population covers only 50% of the phenotype range of the parental lines. If directional epistasis is not a property of the whole genetic architecture, but merely reflects specific interactions between a few loci, data involving more loci and more genetic backgrounds would be expected to reveal less directional epistasis, which seems to be the case here with a striking regularity among the three independent data sources (**Figure 7**).

5. CONCLUDING REMARKS

Unidimensional directional epistasis measures how the properties of genetic architectures change with the phenotype. It has often been confused with scaling. Scale transformation is a common operation in biology, often motivated by the need to make the data suitable for a particular statistical analysis (e.g., enforcing normality). Changing the scale of the phenotype measurement impacts on directional epistasis (Pavlicev et al., 2010), and it is possible to find an arbitrary scale transformation on which directional epistasis becomes negligible (or even is canceled out) in a data set. Applying such *ad hoc* mathematical operations to phenotypes prior to analysis could hardly be considered good practice. First, it has been repeatedly pointed out to biologists that, according to measurement theory, scales do actually have a meaning, and are thus not interchangeable (Wagner et al., 1998; Houle et al., 2011). One of the best examples is fitness, which is essentially multiplicative (Wagner, 2010). Epistasis on fitness thus has to be measured as the deviation from log-linearity, which justifies models of directional epistasis presented above. Obviously, directional epistasis following the power model cancels out on a log scale, but such a double log transformation would be meaningless, and should not be seriously considered. A second reason why scale change does not solve the problem of directional epistasis is that one should not necessarily expect consistent directionality. As exemplified by the chicken example, and illustrated in **Figure 2**, directionality is a local measure of the interlocus curvature of the genotype-phenotype map. It is thus likely that directionality could itself evolve as the phenotype changes (in the presence of third-order epistasis and higher-order interactions, directionality could even change when the phenotype remains constant). Therefore, comparing the properties of genetic architectures across populations or species requires measuring directional epistasis on a common scale.

Recent conceptual and theoretical advances have convincingly demonstrated that what matters in epistasis is not its direct contribution to genetic variation (interaction variance), but rather its propensity to (indirectly) influence the evolution of additive

genetic variance. This propensity can be estimated by looking for specific patterns among epistatic interactions. The directionality of epistasis may be the most obvious, but other patterns are also emerging as candidate contributors to the evolvability of genetic architectures, such as the monotonicity of the genotype-phenotype relationship (closely linked to sign epistasis) (Gjuvsland et al., 2011, 2013), and the robustness or canalization of genetic architectures (Hermisson and Wagner, 2004; Draghi et al., 2010; Fraser and Schadt, 2010; Le Rouzic et al., 2013).

In quantitative genetics and breeding, correctly describing epistasis can improve the prediction of selection responses. In evolutionary genetics, epistasis determines the structure of genetic diversity and variability. At the phylogenetic scale, directional epistasis could contribute to biased anagenesis patterns and affect evolutionary trajectories. Most molecular mechanisms do not simply add up, and the genotype-phenotype relationship has to be curved to some extent. Is the observed curvature (quantified with one or several of the methods described here) consistent with predictions from system-biology models? To what extent is it constrained by the physical properties of the phenotypic trait? Does it vary depending on the trait, on the species? Does it evolve rapidly? The importance of determining directional epistasis for a wide diversity of traits in many organisms has probably been underestimated in the past, but now appears to be a key toward obtaining a better understanding of the general properties of genetic architectures.

ACKNOWLEDGMENTS

I am grateful to Thomas F. Hansen, Estelle Rünneburger, and two reviewers for their careful reading and constructive comments on the manuscript. I acknowledge Paul Siegel, Örjan Carlborg, and Leif Andersson for allowing liberal use of their phenotypic and genetic data on the chicken experiment. Sincere gratitude is expressed to colleagues for advice and discussion, especially José M. Álvarez-Castro. The English text was reviewed by Monika Ghosh.

REFERENCES

Álvarez-Castro, J. M., and Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176, 1151–1167. doi: 10.1534/genetics.106.067348

Álvarez-Castro, J. M., Kopp, M., and Hermisson, J. (2009). Effects of epistasis and the evolution of genetic architecture: exact results for a 2-locus model. *Theor. Popul. Biol.* 75, 109–122. doi: 10.1016/j.tpb.2008.12.003

Álvarez-Castro, J. M., and Le Rouzic, A. (2014). "On the partitioning of genetic variance with epistasis," in *Epistasis and Genetic Architecture*, eds J. Moore and S. Williams (New York, NY: Springer, Humana Press) (in press).

Álvarez-Castro, J. M., Le Rouzic, A., Andersson, L., Siegel, P. B., and Carlborg, Ö. (2012). Modelling of genetic interactions improves prediction of hybrid patterns–a case study in domestic fowl. *Genet. Res.* 94, 255–266. doi: 10.1017/S001667231200047X

Aylor, D. L., and Zeng, Z. B. (2008). From classical genetics to quantitative genetics to systems biology: modeling epistasis. *PLoS Genet.* 4:e1000029. doi: 10.1371/journal.pgen.1000029

Barton, N., and Turelli, M. (2004). Effects of genetic drift on variance components under a general model of epistasis. *Evolution* 58, 2111–2132. doi: 10.1111/j.0014-3820.2004.tb01591.x

Beavis, W. D. (1994). "The power and deceit of QTL experiments: lessons from comparative QTL studies," in *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference* (Chicago, IL), 250–266.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference*, New York, NY: Springer-Verlag LLC.

Butcher, D. (1995). Muller's ratchet, epistasis and mutation effects. *Genetics* 141, 431–437.

Carlborg, Ö., and Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* 5, 618–625. doi: 10.1038/nrg1407

Carlborg, Ö., Hocking, P. M., Burt, D. W., and Haley, C. S. (2004). Simultaneous mapping of epistatic QTL in chickens reveals clusters of QTL pairs with similar genetic effects on growth. *Genet. Res.* 83, 197–209. doi: 10.1017/S0016672304006779

Carlborg, Ö., Jacobsson, L., Ahgren, P., Siegel, P., and Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selection. *Nat. Genet.* 38, 418–420. doi: 10.1038/ng1761

Carlborg, Ö., Kerje, S., Schutz, K., Jacobsson, L., Jensen, P., and Andersson, L. (2003). A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res.* 13, 413–421. doi: 10.1101/gr.528003

Carter, A. J. R., Hermisson, J., and Hansen, T. F. (2005). The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theor. Popul. Biol.* 68, 179–196. doi: 10.1016/j.tpb.2005.05.002

Charlesworth, B. (1990). Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res.* 55, 199–221. doi: 10.1017/S0016672300025532

Cheverud, J. M., and Routman, E. J. (1995). Epistasis and its contribution to genetic variance-components. *Genetics* 139, 1455–1461.

Cheverud, J. M., Vaughn, T. T., Pletscher, L. S., Peripato, A. C., Adams, E. S., Erikson, C. F., et al. (2001). Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. *Mamm. Genome* 12, 3–12. doi: 10.1007/s003350010218

Cockerham, C. C. (1954). An extension of the concept of partitionning herediraty variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39, 859–882.

de Visser, J. A. G. M., Cooper, T. F., and Elena, S. F. (2011). The causes of epistasis. *Proc. Biol. Sci.* 278, 3617–3624. doi: 10.1098/rspb.2011.1537

Demuth, J. P., and Wade, M. J. (2005). On the theoretical and empirical framework for studying genetic interactions within and among species. *Am. Natural.* 165, 524–536. doi: 10.1086/429276

Draghi, J. A., Parsons, T. L., Wagner, G. P., and Plotkin, J. B. (2010). Mutational robustness can facilitate adaptation. *Nature* 463, 353–355. doi: 10.1038/nature08694

Dunnington, E. A., Honaker, C. F., McGilliard, M. L., and Siegel, P. B. (2013). Phenotypic responses of chickens to long-term, bidirectional selection for juvenile body weight — Historical perspective. *Poult. Sci.* 92, 1724–1734. doi: 10.3382/ps.2013-03069

Dunnington, E. A., and Siegel, P. B. (1996). Long-term divergent selection for eight-week body weight in White Plymouth Rock chickens. *Poult. Sci.* 75, 1168–1179. doi: 10.3382/ps.0751168

Elena, S. F., and Lenski, R. E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature* 390, 395–398. doi: 10.1038/37108

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52, 339–433.

Flint, J., and Mackay, T. F. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* 19, 723–733. doi: 10.1101/gr.086660.108

Fraser, H. B., and Schadt, E. E. (2010). The quantitative genetics of phenotypic robustness. *PLoS ONE* 5:e8635. doi: 10.1371/journal.pone.0008635

Gjuvsland, A. B., Hayes, B. J., Omholt, S. W., and Carlborg, Ö. (2007). Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* 175, 411–420. doi: 10.1534/genetics.106.058859

Gjuvsland, A. B., Vik, J. O., Woolliams, J. A., and Omholt, S. W. (2011). Order-preserving principles underlying genotype-phenotype maps ensure high additive proportions of genetic variance. *J. Evol. Biol.* 24, 2269–2279. doi: 10.1111/j.1420-9101.2011.02358.x

Gjuvsland, A. B., Wang, Y., Plahte, E., and Omholt, S. W. (2013). Monotonicity is a key feature of genotype-phenotype maps. *Front. Genet.* 4:216. doi: 10.3389/fgene.2013.00216

Goodnight, C. (1987). On the effect of founder events on epistatic genetic variance. *Evolution* 41, 80–91. doi: 10.2307/2408974

Goodnight, C. (1988). Epistasis and the effect of founder events on the additive genetic variance. *Evolution* 42, 441–454. doi: 10.2307/2409030

Hallander, J., and Waldmann, P. (2007). The effect of non-additive genetic interactions on selection in multi-locus genetic models. *Heredity* 98, 349–359. doi: 10.1038/sj.hdy.6800946

Hansen, T. F. (2013). Why epistasis is important for selection and adaptation. *Evolution* 67, 3501–3511. doi: 10.1111/evo.12214

Hansen, T. F. (2014). "Measuring gene interactions," in *Epistasis and Genetic Architecture*, eds J. Moore and S. Williams (New York, NY: Springer, Humana Press) (in press).

Hansen, T. F., Álvarez-Castro, J. M., Carter, A. J. R., Hermisson, J., and Wagner, G. P. (2006). Evolution of genetic architecture under directional selection. *Evolution* 60, 1523–1536. doi: 10.1111/j.0014-3820.2006.tb00498.x

Hansen, T. F., Pélabon, C., and Houle, D. (2011). Heritability is not evolvability. *Evol. Biol.* 38, 258–277. doi: 10.1007/s11692-011-9127-6

Hansen, T. F., and Wagner, G. P. (2001a). Epistasis and the mutation load: a measurement-theoretical approach. *Genetics* 158, 477–485.

Hansen, T. F., and Wagner, G. P. (2001b). Modeling genetic architecture: a multilinear theory of gene interaction. *Theor. Popul. Biol.* 59, 61–86. doi: 10.1006/tpbi.2000.1508

Hemani, G., Knott, S., and Haley, C. (2013). An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet.* 9:e1003295. doi: 10.1371/journal.pgen.1003295

Hermisson, J., and Wagner, G. P. (2004). The population genetic theory of hidden variation and genetic robustness. *Genetics* 168, 2271–2284. doi: 10.1534/genetics.104.029173

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* 4:e1000008. doi: 10.1371/journal.pgen.1000008

Houle, D. (1992). Comparing evolvability and variability of quantitative traits. *Genetics* 130, 195–204.

Houle, D., Pélabon, C., Wagner, G. P., and Hansen, T. F. (2011). Measurement and meaning in biology. *Q. Rev. Biol.* 86, 3–34. doi: 10.1086/658408

Jarvis, J. P., and Cheverud, J. M. (2009). Epistasis and the evolutionary dynamics of measured genotypic values during simulated serial bottlenecks. *J. Evol. Biol.* 22, 1658–1668. doi: 10.1111/j.1420-9101.2009.01776.x

Jarvis, J. P., and Cheverud, J. M. (2011). Mapping the epistatic network underlying murine reproductive fatpad variation. *Genetics* 187, 597–610. doi: 10.1534/genetics.110.123505

Kao, C. H., and Zeng, Z. B. (2002). Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* 160, 1243–1261.

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proc. R. Soc. Lond. B Biol. Sci.* 143, 102–113. doi: 10.1098/rspb.1954.0056

Kondrashov, F. A., and Kondrashov, A. S. (2001). Multidimensional epistasis and the disadvantage of sex. *Proc. Natl. Acad. Sci. U.S.A.* 98, 12089–12092. doi: 10.1073/pnas.211214298

Kouyos, R. D., Silander, O. K., and Bonhoeffer, S. (2007). Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol. Evol.* 22, 308–315. doi: 10.1016/j.tree.2007.02.014

Lande, R., and Arnold, S. J. (1983). The measument of selection on correlated characters. *Evolution* 37, 1210–1226. doi: 10.2307/2408842

Le Rouzic, A., and Álvarez-Castro, J. M. (2008). Estimation of genetic effects and genotype-phenotype maps. *Evol. Bioinform.* 4, 225–235.

Le Rouzic, A., Álvarez-Castro, J. M., and Carlborg, Ö. (2008). Dissection of the genetic architecture of body weight in chicken reveals the impact of epistasis on domestication traits. *Genetics* 179, 1591–1599. doi: 10.1534/genetics.108.089300

Le Rouzic, A., Álvarez-Castro, J. M., and Hansen, T. F. (2013). The evolution of canalization and evolvability in stable and fluctuating environments. *Evol. Biol.* 40, 317–340. doi: 10.1007/s11692-012-9218-z

Le Rouzic, A., Hansen, T. F., and Houle, D. (2011). A modelling framework for the analysis of artificial selection-response time series. *Genet. Res.* 93, 155–173. doi: 10.1017/S0016672311000024

Le Rouzic, A., Siegel, P. B., and Carlborg, Ö. (2007). Phenotypic evolution from genetic polymorphisms in a radial network architecture. *BMC Biol.* 5:50. doi: 10.1186/1741-7007-5-50

Lenski, R. E., Ofria, C., Collier, T. C., and Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature* 400, 661–664. doi: 10.1038/23245

Liu, G., Dunnington, E. A., and Siegel, P. B. (1994). Responses to long-term divergent selection for eight-week body weight in chickens. *Poult. Sci.* 73, 1642–1650. doi: 10.3382/ps.0731642

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits* (Sinauer Assoc).

Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* 456, 18–21. doi: 10.1038/456018a

Maisnier-Patin, S., Roth, J. R., Fredriksson, Å., Nyström, T., Berg, O. G., and Andersson, D. I. (2005). Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nat. Genet.* 37, 1376–1379. doi: 10.1038/ng1676

Malmberg, R. L., and Mauricio, R. (2005). QTL-based evidence for the role of epistasis in evolution. *Genet. Res.* 86, 89–95. doi: 10.1017/S0016672305007780

Marquez, G., Siegel, P., and Lewis, R. (2010). Genetic diversity and population structure in lines of chickens divergently selected for high and low 8-week body weight *Poult. Sci.* 89, 2580–2588. doi: 10.3382/ps.2010-01034

Omholt, S. W., Plahte, E., Oyehaug, L., and Xiang, K. F. (2000). Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* 155, 969–980.

Otto, S. P. (2007). Unravelling the evolutionary advantage of sex: a commentary on 'mutation–selection balance and the evolutionary advantage of sex and recombination' by Brian Charlesworth. *Genet. Res.* 89, 447–449. doi: 10.1017/S001667230800966X

Otto, S. P., and Feldman, M. W. (1997). Deleterious mutations, variable epistatic interactions, and the evolution of recombination. *Theor. Popul. Biol.* 51, 134–147. doi: 10.1006/tpbi.1997.1301

Pavlicev, M., Le Rouzic, A., Cheverud, J. M., Wagner, G. P., and Hansen, T. F. (2010). Directionality of epistasis in a murine intercross population. *Genetics* 185, 1489–1505. doi: 10.1534/genetics.110.118356

Phillips, P. C. (1998). The language of gene interaction. *Genetics* 149, 1167–1171.

Phillips, P. C. (2008). Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 9, 855–867. doi: 10.1038/nrg2452

Phillips, P. C., Otto, S. P., and Whitlock, M. C. (2000). "Beyond the average: the evolutionary importance of gene interactions and variability of epistatic effects," in *Epistasis and the Evolutionary Process*, eds J. B. Wolf, E. D. Brodie, and M. J. Wade (New York, NY: Oxford University Press), 20–38.

Rönnegård, L., and Valdar, W. (2012). Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability, *BMC Genet.* 13:63. doi: 10.1186/1471-2156-13-63

Shao, H., Burrage, L. C., Sinasac, D. S., Hill, A. E., Ernest, S. R., O'Brien, W., et al. (2008). Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. U.S.A.* 105, 19910–19914. doi: 10.1073/pnas.0810388105

Siegel, P. B. (1962). Double selection experiment for body weight and breast angle at 8 weeks of age in chickens. *Genetics* 47, 1313.

Slatkin, M., and Kirkpatrick, M. (2012). Using known QTLs to detect directional epistatic interactions. *Genet. Res.* 94, 39–48. doi: 10.1017/S0016672312000043

Szathmáry, E. (1993). Do deleterious mutations act synergistically? Metabolic control theory provides a partial answer. *Genetics* 133, 127–132.

Turelli, M., and Barton, N. H. (2006). Will population bottlenecks and multilocus epistasis increase additive genetic variance? *Evolution* 60, 1763–1776. doi: 10.1111/j.0014-3820.2006.tb00521.x

Wade, M. J., Winther, R. G., Agrawal, A. F., and Goodnight, C. J. (2001). Alternative definitions of epistasis: dependence and interaction. *Trends Ecol. Evol.* 16, 498–504. doi: 10.1016/S0169-5347(01)02213-3

Wagner, G. P. (2010). The measurement theory of fitness. *Evolution* 64, 1358–1376.

Wagner, G. P., Laubichler, M. D., and Bagheri, H. C. (1998). Genetic measurement theory of epistatic effects. *Genetica* 102,103, 569–580. doi: 10.1023/A:1017088321094

Wang, T., and Zeng, Z. B. (2006). Models and partition of variance for quantitative trait loci with epistasis and linkage disequilibrium. *BMC Genet* 7:9. doi: 10.1186/1471-2156-7-9

Wilke, C. O., and Adami, C. (2001). Interaction between directional epistasis and average mutational effects. *Proc. R. Soc. Lond. B Biol. Sci.* 268, 1469–1474. doi: 10.1098/rspb.2001.1690

Xu, S. Z. (2003). Theoretical basis of the Beavis effect. *Genetics* 165, 2259–2268.

Yang, R. C. (2004). Epistasis of quantitative trait loci under different gene action models. *Genetics* 167, 1493–1505. doi: 10.1534/genetics.103.020016

Zeng, Z. B., Wang, T., and Zou, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* 169, 1711–1725. doi: 10.1534/genetics.104.035857

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1193–1198. doi: 10.1073/pnas.1119675109

Conflict of Interest Statement: The Guest Associate Editor, Dr. José M. Alvarez-Castro, declares that, despite having collaborated on a publication with the authors in the last 2 years, the review process was handled objectively. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 April 2014; paper pending published: 04 May 2014; accepted: 13 June 2014; published online: 14 July 2014.

*Citation: Le Rouzic A (2014) Estimating directional epistasis. Front. Genet. **5**:198. doi: 10.3389/fgene.2014.00198*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

APPENDIX I: MULTILINEAR EPISTASIS ON A CONTINUOUS GENOTYPE-PHENOTYPE MAP

TWO LOCI

The multilinear model of Hansen and Wagner (2001b) is defined based on a reference genotype, and proposes a change-of-reference operation to recompute the genetic effects in a different genotype, assuming a multilinear genotype-phenotype map. In an arbitrary genotype-phenotype relationship, the multilinear model can be considered to be a local approximation of the multilocus curvature, and epistatic coefficients can be calculated from Taylor polynomial coefficients.

Let $g(y_1, y_2)$ be a continuous and differentiable (at least twice) two-dimensional Genotype-Phenotype function associating a phenotype value P to any genotype combination (y_1, y_2) at two loci. The gradient vector at a particular genotype $\Gamma = (\Gamma_1, \Gamma_2)$ is $\mathbf{D}$ ($D_i = \partial g(y_1, y_2)/\partial y_i|_{\Gamma_1,\Gamma_2}$), and the Hessian matrix is $\mathbf{D}^2$ ($D^2_{i,j} = \partial^2 g(y_1, y_2)/\partial y_i \partial y_j|_{\Gamma_1,\Gamma_2}$). The second-order Tailor series around this genotype Γ is:

$$\begin{aligned} P(y_1, y_2) \simeq\ & g(\Gamma_1, \Gamma_2) + D_1(y_1 - \Gamma_1) + D_2(y_2 - \Gamma_2) \\ & + \frac{1}{2}D^2_{1,1}(y_1 - \Gamma_1)^2 + \frac{1}{2}D^2_{2,2}(y_2 - \Gamma_2)^2 \\ & + D^2_{1,2}(y_1 - \Gamma_1)(y_2 - \Gamma_2). \end{aligned} \tag{A1}$$

Rescaling as $y'_1 = D_1(y_1 - \Gamma_1)$ and $y'_2 = D_2(y_2 - \Gamma_2)$ and neglecting the quadratic terms leads to a multilinear approximation taking the genotype Γ as a reference point:

$$P(y'_1, y'_2) \simeq g(\Gamma_1, \Gamma_2) + y'_1 + y'_2 + y'_1 y'_2 \frac{D^2_{1,2}}{D_1 D_2}, \tag{A2}$$

where it appears clearly that the directionality coefficient of Hansen and Wagner (2001b) is $\varepsilon_{ij} = D^2_{i,j}/D_i D_j$. The quadratic terms $\frac{1}{2}D^2_{1,1}{y'_1}^2$ and $\frac{1}{2}D^2_{2,2}{y'_2}^2$ disappear from the equation as a consequence of the multilinear approximation.

SEVERAL LOCI

The previous approximation can be extended to several loci in a straightforward way:

$$\varepsilon_{ij} = \frac{\partial^2 g}{\partial y_i \partial y_j}\Big|_\Gamma \Big/ \frac{\partial g}{\partial y_i}\Big|_\Gamma \frac{\partial g}{\partial y_j}\Big|_\Gamma. \tag{A3}$$

Developing the third-order Taylor series and neglecting all quadratic terms, the third-order epistatic coefficients can be written as follows:

$$\varepsilon_{ijk} = \frac{\partial^3 g}{\partial y_i \partial y_j \partial y_k}\Big|_\Gamma \Big/ \frac{\partial g}{\partial y_i}\Big|_\Gamma \frac{\partial g}{\partial y_j}\Big|_\Gamma \frac{\partial g}{\partial y_k}\Big|_\Gamma. \tag{A4}$$

The multilinear approximation can thus be easily extended to any number of loci and any order of epistasis, with the n^{th} order epistasis coefficient being the n^{th} mixed partial derivative of the genotype-phenotype function scaled by the product of the first-order derivatives of this function for all loci involved in the interaction.

APPENDIX II: EFFECT OF DIRECTIONAL EPISTASIS ON ARTIFICIAL SELECTION RESPONSE

The impact of directional epistasis on the response to directional selection is rather complex to predict precisely for arbitrary time periods (Carter et al., 2005). Nevertheless, useful approximations can still be derived by making realistic assumptions about the properties of genetic architectures. For instance, Le Rouzic et al. (2011) proposed a model that can be simplified as:

$$\mu_{t+1} = \mu_t + V_{A_t}\beta_t \tag{A5a}$$

$$V_{A_{t+1}} = V_{A_t} + 2\beta_t \varepsilon V^2_{A_t} \tag{A5b}$$

Equation (A5a) is the traditional breeder's equation, formulated as in Lande and Arnold (1983), where V_A is the additive genetic variance, and β the selection gradient, i.e., the slope of the regression between phenotype and relative fitness. Equation (A5b) approximates the impact of directional epistasis on additive variance, summarized by the directionality coefficient ε.

This model requires 3 parameters: μ_0, the initial phenotype, the initial additive variance V_{A_0}, and the epistatic parameter ε. Fitting the model by maximizing its likelihood for phenotype times series including means and variances provide convincing estimates of epistasis, especially when the data include bidirectional artificial selection (Le Rouzic et al., 2011).

Unfortunately, variance time series are not always available from historical data, because either they were measured but not reported in the corresponding publications, or simply because they were not computed, as only the mean phenotype was the center of interest. Moreover, fitting such a complex multidimensional non-linear model can be tricky, and requires significant computer programming input (and possibly having to solve numerical convergence issues). Proposing simpler formulas could therefore be helpful, as they may allow any biologist with basic statistical knowledge to report the strength of directional epistasis based on average phenotype data.

The following calculation is based on several approximations, the main ones being that selection is expected to be constant ($\beta_t = \beta$), and that linkage disequilibrium can be ignored. If directional epistasis is the only phenomenon affecting the selection response, the additive genetic variance is expected to change as in Equation (A5b). Approximating the discrete process by a continuous function leads to the ordinary differential equation $\frac{\mathrm{d}V_A}{\mathrm{d}t} = 2\beta\varepsilon V_A^2$, which can be solved as:

$$V_{A_t} = \frac{V_{A_0}}{1 - 2\beta V_{A_0}\varepsilon t}. \tag{A6}$$

Assuming that directional epistasis is not very strong ($\varepsilon\beta V_{A_0} \ll 1$), the expected phenotype at time t results from the product between the (supposedly constant) selection gradient β and the cumulative change in V_A, which can be calculated as:

$$\mu_t = \mu_0 + \beta \int_0^t V_{A_\tau}\mathrm{d}\tau = \mu_0 - \frac{\log(1 - 2\beta V_{A_0}\varepsilon t)}{2\varepsilon}. \tag{A7}$$

frontiers in
ECOLOGY AND EVOLUTION

HYPOTHESIS AND THEORY ARTICLE
published: 08 September 2014
doi: 10.3389/fevo.2014.00051

Dissecting genetic effects with imprinting

José M. Álvarez-Castro [1,2]*

[1] Department of Genetics, University of Santiago de Compostela, Lugo, Spain
[2] Quantitative Organism Biology, Instituto Gulbenkian de Ciência, Oeiras, Portugal

Edited by:
Rong-Cai Yang, University of Alberta, Canada

Reviewed by:
Bin He, The University of Chicago, USA
Rebekah L. Rogers, University of California, Irvine, USA

***Correspondence:**
José M. Álvarez-Castro, Department of Genetics, Veterinary Faculty, University of Santiago de Compostela, Avda Carvalho Calero, s/n, ES-27002 Lugo, Galiza, Spain
e-mail: jose.alvarez.castro@usc.es

Models of genetic effects are mathematical representations of a genotype-to-phenotype (GP) map that, rather than accounting for a raw map assigning phenotypes to genotypes, rely on parameters with deliberate evolutionary meaning—additive and interaction effects. In this article, the conceptual particularities of genetic imprinting and their implications on models of genetic effects are analyzed. The molecular mechanisms by which imprinted loci affect the relationship between genotypes and phenotypes are known to be singular. Despite its epigenetic nature, the (parent-of-origin-dependent) way in which the alleles of imprinted genes are modified and segregate in each generation is precisely determined, and thus amenable to be represented through conventional models of genetic effects. The Natural and Orthogonal Interactions (NOIA) model framework is here extended to account for imprinting as a tool for a more thorough analysis of the evolutionary implications of this phenomenon. The resulting theory improves and generalizes previous proposals for modeling imprinting.

Keywords: imprinting, individual-referenced models of genetic effects, population-referenced models of genetic effects, NOIA, genetic variance decomposition

INTRODUCTION

Classical models of genetic effects were established almost one century ago for assembling biometric observations with Mendelian genetics (Fisher, 1918; Provine, 1971). This way, mechanistic explanations were provided for interesting properties of quantitative traits that had been revealed in the nineteenth century, particularly the regression toward mediocrity (Galton, 1886). A key concept in this theory is the split of effects of allele substitutions into additive and non-additive components, since the population variance of the additive components was shown to determine the resemblance between relatives within that population (see e.g. Falconer and Mackay, 1996).

The practicality of that rule keeps on being of huge importance nowadays. By assessing the resemblance between relatives for a trait within one generation of a population (which requires tracking relatedness and phenotype scores) it is possible to estimate the additive variance of that trait at that population. That estimate may in its turn be used to predict the resemblance between parents and their offspring and hence the response to selection in the forthcoming generation. Thus, although the theory behind relies on genetic effects, no direct information about the genes underlying a trait in a population is necessary in practice for estimating parameters with convenient predictive power.

With time, molecular, statistical and computational tools have enabled mapping experiments to be performed even in non-model species (see e.g. Rifkin, 2012). The need to update models of genetic effects for making the most of this new source of information was soon pointed out (Cheverud and Routman, 1995), leading to the development of models of genetic effects depicting the GP map as effects of allele substitutions from individual genotypes (Hansen and Wagner, 2001). This is the context in which the Natural and Orthogonal Interactions (NOIA) model of genetic effects was developed (Álvarez-Castro and Carlborg, 2007; Álvarez-Castro and Yang, 2011).

NOIA is a generalization of models of genetic effects that unifies the individual-based formulations mentioned right above with the aforementioned classical approaches, which depict the GP map in terms of effects of allele substitutions averaged over populations. As an example, this approach has enabled analyses of the role of epistatic interactions during the artificial selection process leading to the domestication of chicken (Álvarez-Castro et al., 2008). The classical population-referenced models are convenient for obtaining genetic effects of growth rate from the data generated in quantitative trait loci (QTL) experiments. But, next, those have to be transformed into individual-based genetic effects for analyzing how allele substitutions could have occurred in genes underlying growth rate from the reference of the genotype of the wild ancestors of current domestic chicken. In general, being able to transform between the individual- and the population-referenced approaches opens new opportunities of analyses of gene effects and interactions, as reviewed by Álvarez-Castro (2012).

QTL analyses eventually focussed also on the quest for imprinted genes and the estimation of imprinting effects (Knott et al., 1998). The traditional scheme of either maternal or paternal allele-effect silencing is known not to be universal—the callypige phenotype in sheep being a remarkable counterexample for this (Cockett et al., 1996). Indeed, several alternative patterns of imprinting have been described more recently (e.g. Wolf et al., 2008; Xiao et al., 2013). In general, a gene is imprinted for a trait when heterozygotes with different parent-of-origin of their alleles are associated to different phenotypes. Hence, imprinting always

involves some kind of dominance (since at least one of the two cases will depart from the mid-homozygote expectation).

New models of genetic effects, involving also epistasis, have recently been proposed to detect and analyze imprinted genes (Wolf and Cheverud, 2009). Here, the discussion on how to model genetic effects in the presence of imprinting is resumed with emphasis on the conceptualization (and thus the biological meaning) of all genetic effects involved. Two different options of extending NOIA to imprinting are developed and pondered in order to stress that the meaning of the genetic effects with imprinting must be considered with particular caution.

INDIVIDUAL- AND POPULATION-REFERENCED GENETIC EFFECTS

First, let us recall the most basic expressions and facts of NOIA (from Álvarez-Castro and Carlborg, 2007; Álvarez-Castro et al., 2012). The effects of allele substitutions can be expressed in terms of additive (a) and dominance (d) effects in matrix notation as $\mathbf{G} = \mathbf{SE}$, which, for one non-imprinted locus with two alleles (A_1, A_2) and using the homozygote for the first allele as reference, expands to:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \end{pmatrix} \quad (1)$$

In this expression, $\mathbf{E}$ is the vector of genetic effects (including also the reference point R), $\mathbf{G}$ is the vector of genotypic values (accounting for the expected phenotype for each of the genotypes), and $\mathbf{S}$ is the genetic-effect design matrix, which determines how the genetic effects are defined as a reparameterization of the genotypic values. This point is easier to visualize through the equivalent expression $\mathbf{E} = \mathbf{S}^{-1}\mathbf{G}$:

$$\begin{pmatrix} R \\ a \\ d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -½ & 0 & ½ \\ -½ & 1 & -½ \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} \quad (2)$$

Since $a = (G_{22} - G_{11})/2$ is half the distance between the genotypic values of the two homozigotes, adding two additive effects from the genotypic value of the reference genotype A_1A_1 (G_{11}) brings us to the genotypic value of the other homozygote (G_{22}). Thus, adding one only additive effect brings us to the midpoint between the two homozygotes, from which further adding the dominance effect brings us to the genotypic value of the heterozygote (G_{12}). Indeed, the dominance effect $d = G_{12} - (G_{11} + G_{22})/2$ measures the deviation of the heterozygote from its additive expectation.

More general expressions, enabling the use of any genotype as reference point, have been developed. In any case, the split of effects of allele substitutions from the reference of an individual genotype into additive and interaction components has direct evolutionary meaning. Indeed, assuming that the genotypic values reflect fitness, a quick comparison of the additive and dominance effects provides the equilibrium properties of the system (either one stable or one unstable polymorphic equilibrium, or fixation of a particular allele, which may occur asymptotically with complete dominance). For the simple case of one locus with two alleles, this information can also be retrieved visually from the representation of the raw genotypic values—the genetic effects become more useful for systems of increasing complexity.

On the other hand, the classical additive and interaction population-referenced genetic effects are useful for analyzing properties of particular populations, with given genotype frequencies (p_{ij}, with $p_i = p_{ii} + 1/2p_{12}$ being the allele frequencies and μ the phenotype mean). They are average effects of allele substitutions over populations and they can be obtained by a regression of the genotypic values on the allele content. The general expression for two alleles can be written as:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & -2p_2 & -\frac{p_{12}p_{22}}{2p_1p_2 - ½p_{12}} \\ 1 & p_1 - p_2 & \frac{p_{11}p_{22}}{p_1p_2 - ¼p_{12}} \\ 1 & 2p_1 & -\frac{p_{11}p_{12}}{2p_1p_2 - ½p_{12}} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} \quad (3)$$

The parameters of this model are summarized in **Table 1**. The link between expression (3) and the previous ones comes easy, by just taking into account that the genotypic values remain the same. From any two expressions of this kind, $\mathbf{G} = \mathbf{S_1E_1}$ and $\mathbf{G} = \mathbf{S_2E_2}$, the genetic effects can be transformed into each other directly as:

$$\mathbf{E_2} = (\mathbf{S_2})^{-1}\mathbf{S_1E_1} \quad (4)$$

INTERACTIONS MAKE A DIFFERENCE

Using expression (4), it is easy to derive that a GP map in which $d = 0$ fulfills $\delta = 0$ and $\alpha = a$, regardless of the genotypic frequencies. However, the presence of interactions makes the relationship between individual- and population-referenced

Table 1 | Summary of the parameters of the models in this article.

	All models			Imprinting models
All formulations	G_{11}	G_{12}	G_{22}	G_{21}
	p_{11}	p_{12}	p_{22}	p_{21}
Individual-referenced	R	a	d	d^{12}, d^{21} |i
Population-referenced	μ	α	δ	δ^{12}, δ^{21} |ι

G_{ij} are the genotypic values (expected phenotype of each genotype), with G_{12} for the only heterozygote without imprinting and for one of the two heterozygote options with imprinting (in which case G_{21} stands for the other option). The genotype frequencies (whose subscripts follow the same logic) are p_{ij} and, following the standard notation, the allele frequencies not included in the table are $p_i = p_{ii} + 1/2p_{12}$, $i = 1, 2$. The parameters p_{ij} can also stand as indexes of individual genotypes in the individual-referenced formulation—when one of them equals one and the others equal zero. In the individual-referenced formulation, R stands for the reference point (which is an individual genotype), a for the additive genetic effect and d for the dominance genetic effect. With imprinting, there is an additional imprinting effect, i (in the imprinting-effect model), or two alternative dominance effects, d^{12} and d^{21} (in the two-dominance model; for a justification of the use of the superscripts see Álvarez-Castro and Yang, 2011). In the population-referenced formulation (last row), the corresponding parameters are taken from the Greek alphabet instead of the Latin one (e.g. μ is the population phenotype mean).

genetic effects to be far from trivial—and, indeed, far more interesting (Álvarez-Castro and Le Rouzic, 2014). This is illustrated by two simple examples in **Figure 1**. These graphs show the linear regression (solid line) of the genotypic values (discs) on the allele content (horizontal axis) for a particular population (with specific allele frequencies), as well as the decomposition of the genetic variance (curves) for any allele frequencies.

The first example (**Figure 1A**) shows a case in which the individual-referenced additive genetic effect is nil (the genotypic values of the homozygotes are equal) whereas the dominance effect is not (the genotypic value of the heterozygote is different from them). The slope of the weighted regression of the genotypic values on the allele content provides the population-referenced additive genetic effect, α. In that figure, such regression is shown for a Hardy–Weinberg population with $p_1 = 0.625$, marked with a vertical dashed line. Since the slope of the regression is positive, so it is α. The second example (**Figure 1B**) still shows a case of overdominance (the genotypic values of the homozygotes are lower than the one of the heterozygote, i.e., $d > |a|$), although in this case the individual-referenced additive effect is not nil. However, the regression at $p_1 = 0.625$ has a slope of zero, indicating that this is a (polymorphic) equilibrium point.

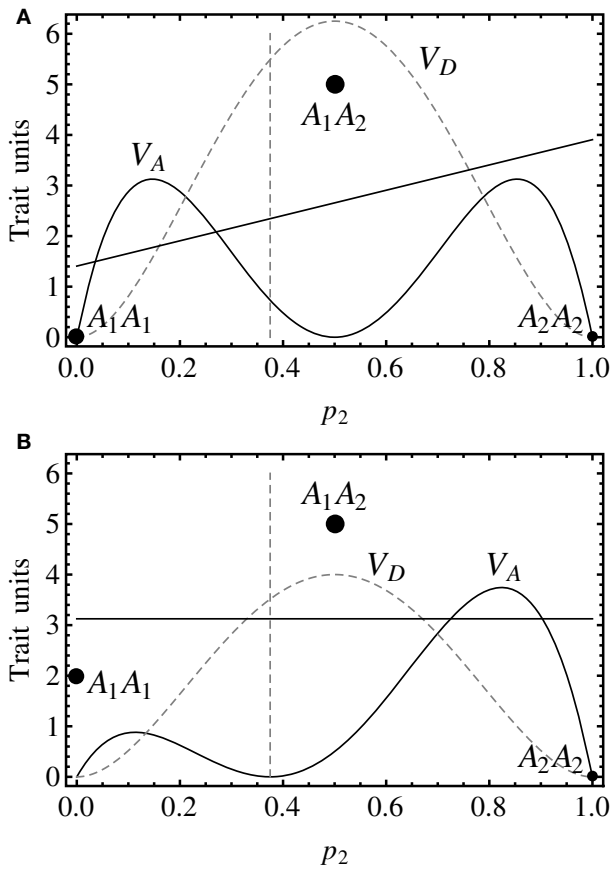


FIGURE 1 | Genotypic values (discs) and variance decomposition (curves) of one-locus, two-allele (A_1 and A_2), non-imprinted genetic systems with overdominance assuming Hardy–Weinberg proportions for all possible allele frequencies (represented by the frequency of A_2, p_2). The variances (black solid curve for additive, gray dashed curve for dominance) are actually plotted as trait units squared. The size of the discs marking the genotypic values are scaled according to $p_1 = 0.625$ (approximately, $p_{11} = 0.14$, $p_{12} = 0.47$, $p_{22} = 0.39$). **(A)** The genotypic values are $G_{11} = 0$, $G_{12} = 5$, $G_{22} = 0$, leading to individual-referenced genetic effects (from the reference of A_1A_1) $a = 0$, $d = 5$. At $p_2 = 0.375$ ($p_1 = 0.625$, marked by the vertical dashed line), the regression of the genotypic values on the proportional allele content (solid line) is an increasing function with slope (and thus population-referenced additive effect) $\alpha = 2.5$, indicating that p_2 would increase under directional selection (toward the equilibrium point, $p_1 = p_2 = 0.5$). **(B)** The genotypic values are the same as in **(A)** but for $G_{11} = 2$, leading to individual referenced genetic effects of $a = -1$, $d = 4$. At $p_2 = 0.375$ ($p_1 = 0.625$, marked by the vertical dashed line), the regression of the genotypic values on the proportional allele content (solid line) has $\alpha = 0$ slope, indicating a polymorphic equilibrium point.

In the context of a population, the decomposition of the genotypic values into additive and interaction effects has its parallel at the level of variances. Indeed, in the second example (**Figure 1B**), the additive variance is nil at $p_1 = 0.625$. Coming back to the first example (**Figure 1A**), the additive variance is not nil at $p_1 = 0.625$ (where the regression slope is not either nil) and, more in general, the additive variance, which determines the selection response, dominates the extremes of the graph (40% of the possible frequencies), indicating very efficient selection response of those populations (toward the equilibrium point, with $p_1 = 0.5$, where the additive variance is nil).

Thus, throughout these examples it becomes evident that interaction makes it possible both to have nil individual-referenced with non-nil population-referenced additive effects and vice versa. Overall, the presence of interactions unveils that individual- and population-referenced genetic effects have different meanings. The later ones reflect properties of populations (the additive effect and the additive variance are nil at equilibrium frequencies) whereas the former ones are effects of allele substitutions from individual references (the additive effect is nil when the homozygotes have equal genotypic values). Keeping this in mind aids interpretation of the subsequent developments and discussion.

MODELING IMPRINTING: HOW MANY ADDITIVE AND DOMINANCE EFFECTS?

When considering one imprinted locus with two alleles, we could be tempted to try to fit it into a one-locus four-allele genetic model, since each of the two alleles (with different nucleotide sequences) may be expressed at the level of the phenotype in two ways (each has two possible methylation stages), thus leading to a total of four variants with potentially different effects on the phenotype. One evident issue coming from this scheme arises when considering how segregation is assumed in a one-locus four-allele model, which does not at all consider transformations of the variants into one another through generations (as it is the case of alleles in imprinted genes). Moreover, even if we dismissed any analyses involving segregation, we could not possibly use the multiallelic model for depicting the differences between phenotypes due to allelic variants, as explained below.

Let the two alleles be A_1 and A_2, just as in the cases without imprinting above. Due to imprinting there now also exist the modified variants $\bar{A}_1$ and $\bar{A}_2$, summing up to a total of four variants as mentioned just above. In a four-allele model of genetic effects, there are six additive effects, three of which can be retrieved from the other three (see e.g. Álvarez-Castro and Yang,

2011). These parameters account for effects of allele substitutions between any possible pair of homozygotes, which in our case would be A_1A_1, A_2A_2, $\bar{A}_1\bar{A}_1$, and $\bar{A}_2\bar{A}_2$. However, none of these genotypes will be present in any of the individuals of our analyses. More to the point, we cannot easily think of those genotypes as putative artificial constructs, since imprinted loci preclude viability under unbalanced dosages of modified alleles (Kono et al., 2004; Kawahara et al., 2007).

Indeed, the two "homozygotes" of our imprinted biallelic locus actually are $A_1\bar{A}_1$ and $A_2\bar{A}_2$—they are allele-wise homozygotes, although not variant-wise homozygotes. Only substitutions implying the pairs A_1-A_2 and $\bar{A}_1$-$\bar{A}_2$ are allowed. Thus, one only additive effect of allele substitutions makes sense in this genetic system, involving substitutions of alleles A_1 and A_2 in each of their variants. In the context of the individual-referenced framework, that effect can be measured in a way analogous to the non-imprinted loci as $a = (G_{22} - G_{11})/2$, just considering that with imprinting the "homozygotes" bear two differently modified allelic variants.

Thus, although properly conceptualizing the additive effects of an imprinted locus may require some reflection, they in the end can be modeled in a way that brings no additional complexity as compared to modeling the non-imprinted case. It is the modeling of the dominance effects that will make the difference. It has been discussed just above that from genotype $A_1\bar{A}_1$ there is one only way of performing two allele substitutions, which leads to genotype $A_2\bar{A}_2$. There are however two possible ways of performing one only allele substitution from that genotype, leading to either $A_1\bar{A}_2$ or $A_2\bar{A}_1$. Consequently, considering two possible dominance effects (one for each parent-of-origin of the two alleles in the heterozygote) emerges as a sensible solution.

To begin with the development of this two-dominance setting, an expression of the genotypic values as a sum of genetic effects of allele substitutions from one reference genotype is firstly provided—as it was done in expression (1) above for a non-imprinted locus. This way (following the same logic as in Álvarez-Castro and Carlborg, 2007; Álvarez-Castro and Yang, 2011), the expression of NOIA from the reference of homozygote $A_1\bar{A}_1$ can be obtained as:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} R \\ a \\ d^{12} \\ d^{21} \end{pmatrix} \tag{5}$$

All parameters are summarized in **Table 1**. The genotypic value of $A_2\bar{A}_2$ is here expressed as the sum of two additive effects from the reference whilst the genotypic values of the heterozygotes involve one additive plus one dominance effect each. The difference between (5) and (1) is that in (5) each heterozygote involves a different dominance effect. By equating the vector of genetic effects in (5) we obtain an extension of expression (2) to imprinting, providing how each of the genetic effects is defined in terms of the genotypic values:

$$\begin{pmatrix} R \\ a \\ d^{12} \\ d^{21} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1/2 \\ -1/2 & 1 & 0 & -1/2 \\ -1/2 & 0 & 1 & -1/2 \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} \tag{6}$$

Thus, for instance, the second dominance effect is defined as $d^{21} = G_{21} - 1/2(G_{11} + G_{22})$. Expression (6) also entails the general individual-referenced formulation of NOIA for one biallelic imprinted locus, by just replacing the first row of the matrix by $(p_{11}, p_{12}, p_{21}, p_{22})$, so that any genotype may be chosen as reference (e.g. $A_2\bar{A}_2$ is the reference when $p_{22} = 1$ and the remaining $p_{ij} = 0$).

For describing the potential response of the imprinted genetic system to one-generation step of selection, a population-referenced formulation [as expression (3) for a non-imprinted locus] is required. Following the same approach as by Álvarez-Castro and Carlborg (2007, Appendix C; see Supplementary Material), such expression can be obtained as:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & -2p_2 & -\dfrac{2p_{12}p_{22}}{(p_{11}+p_{22})(p_{12}+p_{21})} & -\dfrac{2p_{21}p_{22}}{(p_{11}+p_{22})(p_{12}+p_{21})} \\ 1 & p_1 - p_2 & \dfrac{(4p_{11}+p_{21})(p_{11}+p_{22}) - 4p_{11}^2}{(p_{11}+p_{22})(p_{12}+p_{21})} & -\dfrac{p_{21}}{(p_{12}+p_{21})} \\ 1 & p_1 - p_2 & -\dfrac{p_{12}}{(p_{12}+p_{21})} & \dfrac{(4p_{11}+p_{12})(p_{11}+p_{22}) - 4p_{11}^2}{(p_{11}+p_{22})(p_{12}+p_{21})} \\ 1 & 2p_1 & -\dfrac{2p_{11}p_{12}}{(p_{11}+p_{22})(p_{12}+p_{21})} & -\dfrac{2p_{11}p_{21}}{(p_{11}+p_{22})(p_{12}+p_{21})} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \delta^{12} \\ \delta^{21} \end{pmatrix} \tag{7}$$

Using the procedure for inspecting orthogonality of models of genetic effects, also conveyed by Álvarez-Castro and Carlborg (2007, Appendix C; see the Supplementary material), it follows that expression (7) entails an orthogonal decomposition of the genotypic values into additive and dominance components, thus leading to an orthogonal decomposition of the genetic variance. The two dominance effects are however not orthogonal to each other. Overall, it is possible to model a biallelic imprinted locus using one additive and two dominance genetic effects, which makes it straightforward to keep track of the biological meaning of the parameters, in analogy with the non-imprinted case.

IMPRINTING AS A GENETIC EFFECT

The previous setting can be used for detecting imprinting by just developing a procedure for testing whether the two dominance effects are significantly different. To this aim, it seems however more convenient to design a model in which a parameter accounts

for the difference between the two heterozygotes, thus leading to a more direct test for imprinting—consisting in just checking whether that parameter is significantly different from zero. Actually, this is in general terms the approach commonly chosen to model imprinting (see e.g. Wolf et al., 2008). Hereafter, NOIA is extended following that approach and thus implemented with a parameter to account for the putative difference between the heterozygotes with different parent-of-origin. As in the previous section, an expression of effects of allele substitutions from the reference of homozygote $A_1\bar{A}_1$ is here provided in the first place, as:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \\ i \end{pmatrix} \tag{8}$$

This model is designed for using the midpoint between the two heterozygotes to define the dominance effect and the deviations of the two heterozygotes from that point as the imprinting effect. A graphical comparison explaining how the three models shown in this article (the non-imprinted model, the two-dominances model and the imprinting-effect model) decompose the genotypic values is shown in **Figure 2**. By equating the vector of genetic effects in (8) it follows:

$$\begin{pmatrix} R \\ a \\ d \\ i \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1/2 \\ -1/2 & 1/2 & 1/2 & -1/2 \\ 0 & -1/2 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} \tag{9}$$

From this expression it immediately follows that indeed $d = 1/2(G_{12} + G_{21}) - 1/2(G_{11} + G_{22})$ (i.e., the dominance effect measures the distance of the midpoint between the two heterozygotes and the additive expectation) and $i = 1/2(G_{21} - G_{12})$ (i.e., the imprinting effect measures the distance of the heterozygotes from the midpoint between them). Expression (9) provides a general individual-referenced formulation, analogously to (6) for the two-dominances model in the previous section. Also in an analogous way as in that section, an orthogonal population-referenced formulation of the imprinting-effect model can be obtained as:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & -2p_2 & -\frac{p_{22}(p_{12}+p_{21})}{2p_1p_2 - 1/2(p_{12}+p_{21})} & 0 \\ 1 & p_1 - p_2 & \frac{p_{11}p_{22}}{p_1p_2 - 1/2(p_{12}+p_{21})} & \frac{-2p_{21}}{p_{12}+p_{21}} \\ 1 & p_1 - p_2 & \frac{p_{11}p_{22}}{p_1p_2 - 1/2(p_{12}+p_{21})} & \frac{2p_{12}}{p_{12}+p_{21}} \\ 1 & 2p_1 & -\frac{p_{11}(p_{12}+p_{21})}{2p_1p_2 - 1/2(p_{12}+p_{21})} & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \delta \\ \iota \end{pmatrix} \tag{10}$$

In this case, the three genetic (additive, dominance and imprinting) effects are fully orthogonal. The independence of the parameters makes this expression to resemble expression (3). Indeed, the decomposition of the genotypic values of the homozygotes into additive and dominance effects in (3) holds in (10), since p_{12} in (3) is equivalent to $(p_{12} + p_{21})$ in (10). Concerning the heterozygotes, in the imprinted case we have two instead of one, leading to an extra row in the genetic-effects design matrix in (10), and there is an extra (imprinting) term in the decomposition, coming from the fourth column of that matrix. That term actually makes the only difference of the decomposition of the genetic effects of the heterozygotes as compared with the decomposition of the heterozygote in the non-imprinted case (3).

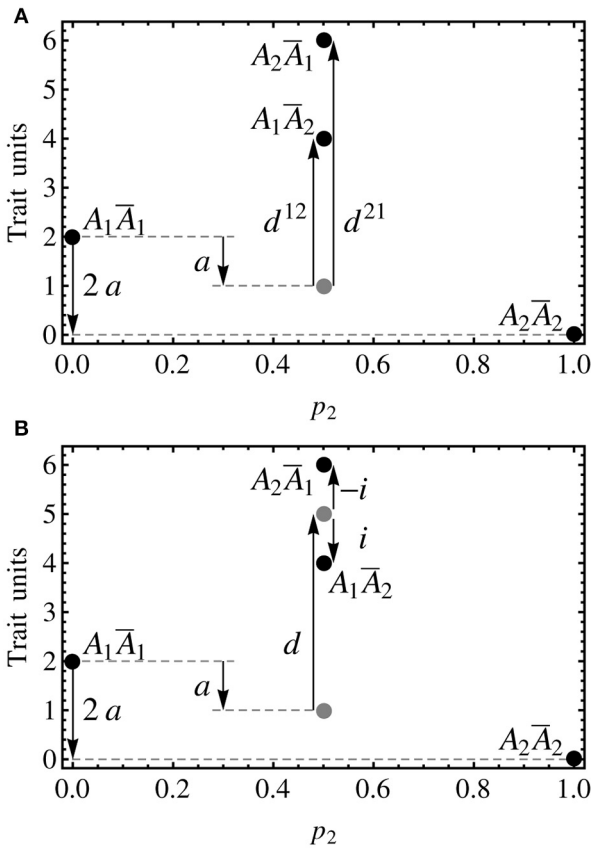


FIGURE 2 | Individual-referenced genetic effects proposed in the text for a one-locus, two-allele, imprinted genetic system. As in the previous figure, the alleles are A_1 and A_2, with the variants due to imprinting being $\bar{A}_1$ and $\bar{A}_2$. Although no population frequencies are considered in this figure, the notation of the axes is kept consistent with the other figures, with $p_2 = 0.5$ indicating the heterozygotes. Also the genotypic values (black discs) are mostly kept, with the one of the heterozygote of **Figure 1** (A_1A_2, $G_{12} = 5$) being the midpoint between the ones of the two heterozygotes in this figure ($A_1\bar{A}_2$ and $A_2\bar{A}_1$, $G_{12} = 4$ and $G_{21} = 6$, respectively). **(A)** The two-dominances model is a natural extension of the non-imprinted case that consists in introducing two dominance parameters, one for each heterozygote. As well as in the non-imprinted case, the dominance effects measure departures of the heterozygotes from their additive expectation (gray disc). **(B)** The imprinting-effect model keeps one only dominance effect that accounts for the departure of the midpoint between the two heterozygotes (upper gray disc) and the midpoint between the homozygotes (lower gray disc), and adds up an imprinting effect for accounting for the distance between the heterozygotes and their midpoint (upper gray disc). Thus, the dominance effect in this model coincides with that of a non-imprinted model (i.e., the one that would be obtained if imprinting was just disregarded).

VARIANCE DECOMPOSITION WITH IMPRINTING

The previous expressions and arguments can be extended to the decomposition of the genetic variance with an imprinting

variance component, which can easily be obtained from the model in matrix notation above (10) by following the formulae provided by Álvarez-Castro and Yang (2011). In expressions (12) and (13) of that article, the additive and the dominance variance have been obtained as $V_A = \mathrm{P}_G^{\mathrm{T}}(\alpha_G \circ \alpha_G)$ and $V_D = \mathrm{P}_G^{\mathrm{T}}(\delta_G \circ \delta_G)$, respectively. In an analogous way (by means of analogous intermediate definitions; see Supplementary Material), a general expression for the imprinting variance can be provided simply as:

$$V_O = \mathbf{P}_G^{\mathrm{T}}(\boldsymbol{\iota}_G \circ \boldsymbol{\iota}_G) \tag{11}$$

Since V_I traditionally stands for the epistatic variance, the subscript O is here chosen for the imprinting variance, ultimately coming from a differential effect of the alleles depending on their parent-of-origin. In any case, it is also possible to obtain the decomposition of the genetic variance by getting all three variance components at the same time, by just following expressions (14) and (15) in Álvarez-Castro and Yang (2011). Indeed, the imprinting variance component emerges from that formulae as a new term due to feeding them with expression (10).

By obtaining the variance decomposition in any of the ways described above (each individually or all simultaneously), it is easy to check that the additive and the dominance variances actually remain the same as for a non-imprinted biallelic locus. Assuming for simplicity the Hardy–Weinberg proportions, they are $V_A = 2p_1p_2[a + d(p_1 - p_2)]^2$, $V_D = (2dp_1p_2)^2$ (see e.g. Falconer and Mackay, 1996)—, whilst the imprinting variance component can be expressed simply as:

$$V_O = 2i^2 p_1 p_2 \tag{12}$$

Figure 3 shows the decomposition of the genetic variance for two cases of imprinting. The genotypic values in **Figure 3A** are the same as in **Figure 2**, and thus they also fit the non-imprinted case in **Figure 1B**, in which the genotypic value of the heterozygote (A_1A_2, $G_{12} = 5$) is the midpoint between the genotypic values of the two heterozygotes in **Figure 3A** ($A_1\bar{A}_2$ and $A_2\bar{A}_1$, $G_{12} = 4$ and $G_{21} = 6$, respectively). Therefore, the additive effects coincide in both cases and the dominance value of the imprinting-effect model in **Figure 3A** coincides with the simpler non-imprinted model in **Figure 1B**. Hence, the additive and the dominance variances coincide in both graphs. In **Figure 3A** there is, though, an extra (imprinting) term of the genetic variance decomposition.

As it is the case for dominance, the imprinting variance is higher for intermediate frequencies. In **Figure 3A**, the relatively small imprinting effect (relatively short distance between the two heterozygotes) leads to a small imprinting variance for all allele frequencies. In **Figure 3B**, however, it is shown that with larger differences between the two heterozygotes the imprinting variance may dominate the variance decomposition at almost any allele frequencies. And this actually occurs in practice, since this case fits to the callypige pattern mentioned above (with equal or similar phenotype values of the two homozygotes and one of the heterozygotes, relative to a higher value of the remaining heterozygote). Imprinting is thus—as well as other allele interactions (Álvarez-Castro and Le Rouzic, 2014)—a phenomenon that may by itself condition little responses to selection in the face of high genetic variances.

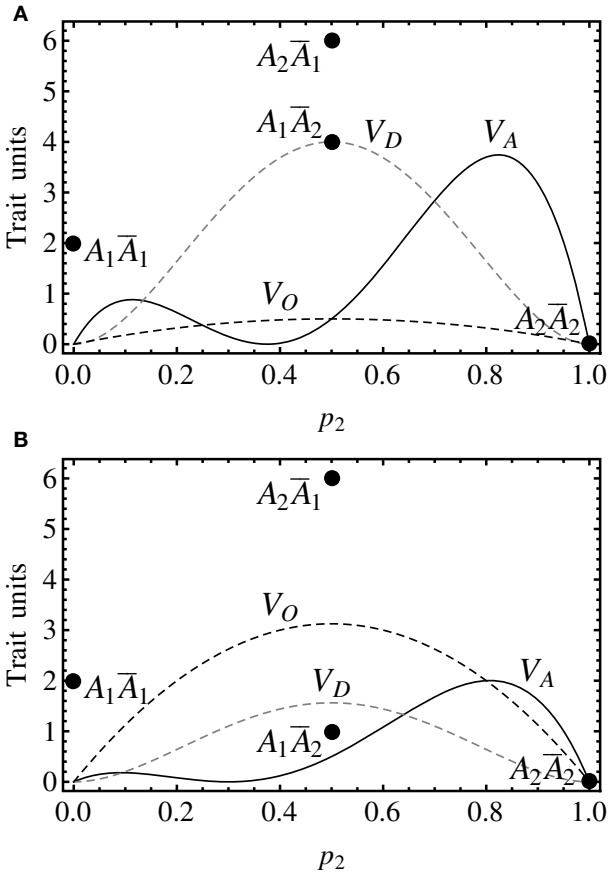


FIGURE 3 | Genotypic values (discs) and variance decomposition (curves) of one-locus, two-allele, imprinted genetic systems assuming Hardy–Weinberg proportions. The notation is in accordance with the previous figures (with the addition of a black dashed curve for the imprinting variance). **(A)** The genotypic values are $G_{11} = 2$, $G_{12} = 4$, $G_{21} = 6$, $G_{22} = 0$, leading to individual-referenced genetic effects (from the reference of $A_1\bar{A}_1$) $a = -1$, $d = 4$, $i = -1$. Since the additive and the dominance effects are the same as in **Figure 1B**, the additive and the dominance variances coincide and the equilibrium point also remains at $p_2 = 0.375$. **(B)** The genotypic values are the same as in **(A)** but for $G_{21} = 1$ (at the midpoint between the two homozygotes), leading to individual referenced genetic effects of $a = -1$, $d = 2.5$, $i = -2.5$. The equilibrium point occurs here at $p_2 = 0.3$.

Incidentally, this particular claim could not be supported using the two-dominances model alone. Indeed, that model does not provide a separate term accounting for the variance explained by the difference between the two heterozygotes. Instead, it leads to a dominance variance that is different from the one of this imprinting-effect model (and thus also from the one of the non-imprinted case), and it actually equals the sum of the classical dominance variance V_D and the imprinting variance V_O as expressed above (11, 12).

COMPARISONS TO PREVIOUS MODELS

Xiao et al. (2013) have recently proposed a model of imprinting based on the (non-imprinted) NOIA model. They take the option of implementing an explicit imprinting parameter, which in their mathematical construction is closely related to the additive effect, rather than to the dominance effect as in the imprinting-effect model developed above (8–10). Since it is in this article acknowledged that modeling imprinting requires some improvisation as compared to other facts of genetic architecture, several different solutions could be possible—it is not intended here to pose any objective criticism on that choice by itself.

The developments by Xiao et al. (2013) are indeed inspired in the NOIA model and they provide both statistical (i.e., population-referenced) and functional (which are not population-referenced) formulations. However, their models are difficult to be considered as pure extensions of the NOIA model. A very simple counterexample for this can be shown through their expression (12), from which it follows that they define the functional additive effect as $r_1 = G_{22} - G_{11}$, whereas in the NOIA model it is defined as $a = (G_{22} - G_{11})/2$. This can be easily derived e.g. from (2) for the non-imprinted case, and also from (6) and (9) for the extensions to imprinting provided in this article.

Xiao et al. (2013) carried out simulations to prove that their statistical models are more appropriate (due to orthogonality) for detecting allelic effects than their functional developments. This effort seems to be rather futile since the functional formulations are in general not developed with that motivation in mind, but mainly for representing the GP map as effects of allele substitutions from individual references (Hansen and Wagner, 2001; Álvarez-Castro and Carlborg, 2007; Álvarez-Castro, 2012; and also summarized above). In any case, the statistical models of imprinting by Xiao et al. (2013) are admittedly not fully orthogonal as the imprinting-effect model provided above (10), but only under certain conditions e.g. (but not only) under the Hardy–Weinberg proportions.

Wolf and Cheverud (2009, Appendix 2) had also provided a model with an explicit imprinting parameter that is orthogonal under the Hardy–Weinberg proportions. As well as Xiao et al. (2013), they make the point that, also with imprinting, extensions to multiple loci with epistasis come naturally using the Kronecker product of genetic-effect design matrices (following Tiwari and Elston, 1997), which incidentally applies directly also to the models of imprinting provided in this article. However, Wolf and Cheverud (2009) do not provide explicit expressions for performing variance decompositions.

Neither they discuss an explicit link of their statistical setting to a functional formulation, although their expressions (4) and (5) fit to an extension of the physiological model (Cheverud and Routman, 1995, which is an alternative to statistical formulations with the unweighted population mean as reference point) rather than to the F_2 model they initially follow in their developments. More to the point, in their previous work on imprinting (Wolf et al., 2008) they made an extension of the F_∞ model, another alternative to the classical statistical formulations.

There is also a previous work in which a two-dominance strategy has been chosen to model imprinting, by Santure and Spencer (2011). They have adapted several standard quantitative approaches to derive quantitative genetics parameters in the presence of imprinting, which is implemented as in this article, in the form of one dominance effect for each heterozygote. The different approaches considered in that article lead to different results, but none of them enables an orthogonal decomposition of the genetic variance into additive and dominance (due to the two dominance effects) components. For several of those approaches, expressions of the covariances due to lack of orthogonality could not be derived.

DISCUSSION

Since models of genetic effects are mathematical expressions aimed to enable the estimation of parameters with particular biological interpretations, their development is often directed to a predefined target. The difficulties of these developments often consist in reaching the mathematical properties that are in accordance with the desired biological meanings. With imprinting, there appears an extra layer of issues to be solved, ultimately coming from the fact that many combinations of alleles or allele variants will never occur (not even artificially). For solving that issue, modeling that $A_2\bar{A}_2$ can be reached by performing two equal allele substitutions from $A_1\bar{A}_1$ entails a very sensible and practical solution (even acknowledging that this is not in reality the case).

Standing from this point, and facing the presence of two different heterozygotes (and their genotypic values), it appears natural to think of accounting for two different dominance effects, analogous to the one dominance effect in the non-imprinted case. This solution, here called the two-dominances model, is not only feasible but, as shown in **Figure 2A**, rather clean by construction. It indeed leads naturally to an orthogonal variance partition into additive and interaction components. However, with this setting it may not be completely straightforward to detach imprinting as an effect either to test or to analyze in terms of evolutionary properties.

Traditional models of imprinting have embraced the option of implementing an explicit imprinting effect, which is here called the imprinting-effect model. Dominance is modeled as a departure from an additive (non-dominance) expectation. For modeling imprinting in an analogous way, a non-imprinting reference has to be considered. Due to the particularities of imprinting, this reference has to be a construct. Indeed, as explained above, we cannot just remove imprinting effects from our alleles and expect that the resulting genotypes exist or could even be viable, and there seems to be no biological justification for choosing one of the heterozygotes as the non-imprinted reference against the other one. Hence, the midway between the two of them is in this article set as a non-imprinted fictitious reference. In **Figure 2B** it can be seen that this leads for instance to a definition of the dominance effect in terms of points (gray discs) that are not genotypic values (black discs). In any case, several advantages come from this choice.

The imprinting-effect model here provided leads to a fully orthogonal setting, which entails a clear advantage over previous models. This is optimal in the first place for testing for statistical significance of the imprinting parameter. Furthermore, this

setting can be described as a pure extension of a non-imprinting case with the heterozygote at the midpoint between the two imprinted heterozygote options. The variance partition, in particular, remains equal to the non-imprinting case in what regards all variance components except from the imprinting variance, which is of course absent in the non-imprinting case. This enables extremely convenient comparisons: the equilibrium points of the two cases will be the same, with a slowed down speed of phenotype change along generations for the imprinted case, which shall be more noticeable for increasing proportions of the imprinting variance component in the genetic variance partition (since the proportion of the additive component of the phenotypic variance decreases accordingly).

Besides population-referenced orthogonal expressions, individual-based formulations are in this article provided. When using any expressions in this article, the choice of a formulation and a reference point must be based on the mathematical properties and/or biological meaning that fits the particular question to be addressed. Each choice leads to different numerical values of at least some of the parameters in an applied case and thus not paying enough attention to picking the correct expression may be misleading. An illustration of such requisite of awareness on the specific kind of genetic effects used in each case follows.

In their article on imprinting and epistasis, Wolf and Cheverud (2009) claim, based on a previous work (Cheverud, 2000), that "additive-by-dominance indicates that the additive effect of the first locus depends on (i.e., changes as a function of) the genotype present in the second locus, while the dominance effect of the second locus depends on the genotype present at the first locus." This is true when analyzing a genetic system with the physiological model (that is, for physiological additive-by-dominance genetic effects). Functional formulations are meant to express genetic effects from the reference of individual genotypes, i.e., as individual-based formulations. Mathematically, it is straightforward to use those expressions also from other reference points and, when doing so, it can be shown that they then coincide with statistical (population-referenced) formulations under certain conditions [Álvarez-Castro and Carlborg, 2007, expression (7)]. Both the F_∞, the F_2 and the physiological models are instances of this situation: they thus may fit both to functional and to statistical interpretations and this is why the afore-cited sentence holds true within its particular context.

However, it is worthwhile noting that the referred sentence is not true for additive-by-dominance genetic effects of any model or formulation, and in particular it cannot be applied if the genetic effects are orthogonal (in the context a population under study) and conditions (7) of Álvarez-Castro and Carlborg (2007) do not hold. Indeed, in those instances it may well be that dominance-by-dominance interactions generate statistical additive-by-dominance interaction at genetic systems for which the latest equals zero under the physiological model. Such a phenomenon is analogous to the simpler instance shown in **Figure 1**, where the presence of dominance interaction is shown to generate additive variance in a genetic system where there are no difference between the homozygotes (i.e., nil functional additive effects). Interestingly, this hierarchical behavior works in a different way when it comes to imprinting. Indeed, the imprinting-effect model developed above is structured such that functional imprinting alone (with neither functional dominance nor functional additive effects) generates neither dominance nor additive variance, as it can be seen by the fact that these variances do not depend on the imprinting effect.

Overall, it is in general crucial to mind the biological meaning of the models in order to make the choice of the particular expression to be used in each particular case. In relation with this, NOIA conveniently provides expressions that work as a change-of-reference tool so that the genetic effects required to a particular question can be obtained from any others. The scope of that tool applies to transformations between the two-dominances and the imprinting-effect models developed above, which differ in the presence/absence of an explicit genetic imprinting effect. The choices of formulations are therefore not excluding, but potentially informative about different aspects in the analysis of a particular situation under study as long as the resulting values of the genetic effects (or variance decompositions) are interpreted in the light of the particular form of the genetic model used.

This article stands on recent advances in genetic modeling for carrying out new theoretical developments to the aid of the analysis of genetic imprinting. The models here developed improve previous proposals by providing both functional and statistical formulations that enable an orthogonal partition of the genotypic values and the genetic variance with a separate component for imprinting, which enables both better estimation of, and insight on, imprinted genes. Besides, imprinting may here be conceived also as an excuse or a challenge in order to elaborate on the logics behind the development of models of genetic effects—what are they intended for, which difficulties condition their stage of development, how to face them. Overall, one more step in the generalization of models of genetic effects is here provided, as well as keys about the way models of genetic effects may keep on being developed.

ACKNOWLEDGMENT

The author acknowledges the two reviewers for their suggestions, which improved this manuscript. The Autonomous Administration Xunta de Galicia provided funding for this research through project EM2014/024.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fevo.2014.00051/abstract

REFERENCES

Álvarez-Castro, J. M., Carlborg, O., and Ronnegard, L. (2012). "Estimation and interpretation of genetic effects with epistasis using the NOIA model," in *Quantitative Trait Loci (QTL): Methods and Protocols*, ed S. Rifkin (New York, NY: Springer, Humana Press), 191–204.

Álvarez-Castro, J. M. (2012). Current applications of models of genetic effects with interactions across the genome. *Curr. Genomics* 13, 163–175. doi: 10.2174/138920212799860689

Álvarez-Castro, J. M., and Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176, 1151–1167. doi: 10.1534/genetics.106.067348

Álvarez-Castro, J. M., and Le Rouzic, A. (2014). "On the partitioning of genetic variance with epistasis," in *Epistasis: Methods and Protocols,* eds J. H. Moore and S. M. Williams (New York, NY: Springer, Humana Press). (in press).

Álvarez-Castro, J. M., Le Rouzic, A., and Carlborg, Ö. (2008). How to perform meaningful estimates of genetic effects. *PLoS Genet.* 4:e1000062. doi: 10.1371/journal.pgen.1000062

Álvarez-Castro, J. M., and Yang, R.-C. (2011). Multiallelic models of genetic effects and variance decomposition in non-equilibrium populations. *Genetica* 139, 1119–1134. doi: 10.1007/s10709-011-9614-9

Cheverud, J. M. (2000). "Detecting epistasis among quantitative trait loci," in *Epistasis and the Evolutionary Process*, eds J. B. Wolf, E. D. Brodie and M. J. Wade (Oxford: Oxford University Press), 58–81.

Cheverud, J. M., and Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics* 139, 1455–1461.

Cockett, N. E., Jackson, S. P., Shay, T. L., Farnir, F., Berghmans, S., Snowder, G. D., et al. (1996). Polar overdominance at the ovine callipyge locus. *Science* 273, 236–238. doi: 10.1126/science.273.5272.236

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Harlow: Prentice Hall.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52, 339–433.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Anthrop. Inst. Great Br. Ireland* 15, 246–263.

Hansen, T. F., and Wagner, G. P. (2001). Modeling genetic architecture: a multilinear theory of gene interaction. *Theor. Popul. Biol.* 59, 61–86. doi: 10.1006/tpbi.2000.1508

Kawahara, M., Wu, Q., Takahashi, N., Morita, S., Yamada, K., Ito, M., et al. (2007). High-frequency generation of viable mice from engineered bi-maternal embryos. *Nat. Biotechnol.* 25, 1045–1050. doi: 10.1038/nbt1331

Knott, S. A., Marklund, L., Haley, C. S., Andersson, K., Davies, W., Ellegren, H., et al. (1998). Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* 149, 1069–1080.

Kono, T., Obata, Y., Wu, Q., Niwa, K., Ono, Y., Yamamoto, Y., et al. (2004). Birth of parthenogenetic mice that can develop to adulthood. *Nature* 428, 860–864. doi: 10.1038/nature02402

Provine, W. B. (1971). *The Origins of Theoretical Population Genetics*. Chicago, IL: University of Chicago Press.

Rifkin, S. A. (2012). *Quantitative Trait Loci (QTL)*. New York, NY: Springer. doi: 10.1007/978-1-61779-785-9

Santure, A. W., and Spencer, H. G. (2011). Quantitative genetics of genomic imprinting: a comparison of simple variance derivations, the effects of inbreeding, and response to selection. *G3 (Bethesda)* 1, 131–142. doi: 10.1534/g3.111.000042

Tiwari, H. K., and Elston, R. C. (1997). Deriving components of genetic variance for multilocus models. *Genet. Epidemiol*. 14, 1131–1136.

Wolf, J. B., and Cheverud, J. M. (2009). A framework for detecting and characterizing genetic background-dependent imprinting effects. *Mamm. Genome* 20, 681–698. doi: 10.1007/s00335-009-9209-2

Wolf, J. B., Cheverud, J. M., Roseman, C., and Hager, R. (2008). Genome-wide analysis reveals a complex pattern of genomic imprinting in mice. *PLoS Genet* 4:e1000091. doi: 10.1371/journal.pgen.1000091

Xiao, F., Ma, J., and Amos, C. I. (2013). A unified framework integrating parent-of-origin effects for association study. *PLoS ONE* 8:e72208. doi: 10.1371/journal.pone.0072208

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 05 August 2014; published online: 08 September 2014.

*Citation: Álvarez-Castro JM (2014) Dissecting genetic effects with imprinting. Front. Ecol. Evol. **2**:51. doi: 10.3389/fevo.2014.00051*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution.

Appendix

Regression approach for developing orthogonal genetic effects

This approach consists in computing the genetic effects from the regression of genotypic values to the allele content, as Fisher (1918) proposed (see e.g. Falconer and Mackay (1996)). Álvarez-Castro and Carlborg (2007) expressed the genotypic values as $G(N)=E(G)+\beta N$, where N stands for the number of A_2 alleles. The intercept of the regression, $E(G)$, is the expectation of the genotypic values and the regression coefficient is $\beta = \mathrm{Cov}(G,N)/\mathrm{Var}(N)$. The additive effects come from the linear regression itself, whereas the interaction terms come from the departures—the distances between the regression and the original genotypic values (for further details, see Álvarez-Castro and Carlborg 2007). In the two-dominance model, each heterozygote determines one dominance effect (as represented in Figure 2A), whereas in the imprinting-effect model, the dominance effect is defined as in the non-imprinting case by taking the midpoint between the two heterozygotes as the one heterozygote required for making that definition. The imprinting effect is defined afterwards from the departures of the real heterozygotes from that midpoint (as represented in Figure 2B).

A genetic-effect design matrix, **S**, is orthogonal for a set of genotype frequencies when $\mathbf{S}^{\mathrm{T}}\mathbf{DS}$ is diagonal, with $\mathbf{D}=\mathrm{Diag}[(p_{ij})]$ *i.e.* the diagonal matrix with the genotypic frequencies at its diagonal.

Variance components from the decomposition of the genotypic values

Following Álvarez-Castro and Yang (2011), the decomposition of genotypic values into additive and interaction terms that is implicit in an expression of the type $\mathbf{G}=\mathbf{SE}$ can be made explicit as $\mathbf{G}_{dec}=\mathbf{S}\mathrm{Diag}[\mathbf{E}]$. From here, the decomposition of the genetic variance takes the form of a vector (with the variance components) by just computing $\mathbf{V}=\mathbf{P}_G^{\mathrm{T}}\left(\mathbf{G}_{dec}\circ\mathbf{G}_{dec}\right)$, with $\mathbf{P}_G^{\mathrm{T}}=(p_{ij})$. For computing the imprinting variance separately in the context of a one-locus two-allele model, the imprinting vector can be defined as $\boldsymbol{\iota}_G=\mathbf{S}\mathrm{Diag}[(0,0,0,1)]\mathbf{E}$, to then apply expression (11).

frontiers in GENETICS

GENERAL COMMENTARY
published: 08 December 2014
doi: 10.3389/fgene.2014.00427

Corrigendum for "Dissecting genetic effects with imprinting"

José M. Álvarez-Castro*

Department of Genetics, Universidade de Santiago de Compostela, Lugo, Spain
**Correspondence: jose.alvarez.castro@usc.es*

Edited and reviewed by:
Rong-Cai Yang, University of Alberta, Canada

Keywords: imprinting, NOIA, individual-referenced models of genetic effects, population-referenced models of genetic effects, genetic variance decomposition

A corrigendum on

Dissecting genetic effects with imprinting
by Álvarez-Castro, J. M. (2014). Front. Ecol. Evol. 2:51. doi: 10.3389/fevo.2014.00051

Corrigendum:
In the article of the Frontiers Research Topic Issue on Models and Estimation of Genetic Effects "Dissecting genetic effects with imprinting," by Álvarez-Castro (2014), the citation of the work in press by Álvarez-Castro and Le Rouzic is no longer correct since its publication has in the end been postponed to 2015 (Álvarez-Castro and Le Rouzic, 2015). The same holds for the citation of that reference in press in the article "Estimating directional epistasis," by Le Rouzic (2014), also within the Frontiers Research Topic Issue on Models and Estimation of Genetic Effects.

REFERENCES

Álvarez-Castro, J. M. (2014). Dissecting genetic effects with imprinting. *Front. Ecol. Evol.* 2:51. doi: 10.3389/fevo.2014.00051

Álvarez-Castro, J. M., and Le Rouzic, A. (2015). "On the partitioning of genetic variance with epistasis," in *Epistasis: Methods and Protocols*, eds J. H. Moore and S. M. Williams (New York, NY: Springer, Humana Press) (in press).

Le Rouzic, A. (2014). Estimating directional epistasis. *Front. Genet.* 5:198. doi: 10.3389/fgene.2014.00198

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 November 2014; accepted: 20 November 2014; published online: 08 December 2014.
*Citation: Álvarez-Castro JM (2014) Corrigendum for "Dissecting genetic effects with imprinting." Front. Genet. **5**:427. doi: 10.3389/fgene.2014.00427*
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 09 March 2012
doi: 10.3389/fgene.2012.00030

Clarifying the relationship between average excesses and average effects of allele substitutions

José M. Álvarez-Castro[1]* and Rong-Cai Yang[2,3]

[1] *Department of Genetics, University of Santiago de Compostela, Lugo, Spain*
[2] *Research and Innovation Division, Alberta Agriculture and Rural Development, Edmonton, Alberta, AB, Canada*
[3] *Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, AB, Canada*

Edited by:
Jason Wolf, University of Bath, UK

Reviewed by:
Chen-Hung Kao, Academia Sinica, Taiwan
Alan Templeton, Washington University, US Minor Outlying Islands

****Correspondence:***
José M. Álvarez-Castro, Department of Genetics, Veterinary Faculty, University of Santiago de Compostela, Avda Carvalho Calero, s/n, ES-27002 Lugo, Galiza, Spain.
e-mail: jose.alvarez.castro@usc.es

Fisher's concepts of average effects and average excesses are at the core of the quantitative genetics theory. Their meaning and relationship have regularly been discussed and clarified. Here we develop a generalized set of one locus two-allele orthogonal contrasts for average excesses and average effects, based on the concept of the effective gene content of alleles. Our developments help understand the average excesses of alleles for the biallelic case. We dissect how average excesses relate to the average effects and to the decomposition of the genetic variance.

Keywords: average effects, average excesses, effective gene content, models of genetic effects, non-equilibrium populations

INTRODUCTION

Since Fisher (1918), partitioning of the genotypic values at a locus into additive and dominance effects has been used for conventional quantitative genetic analyses and recently for mapping quantitative trait loci (QTL; see, e.g., Lynch and Walsh, 1998). Numerous statistical models have been proposed for such partitioning. Some of them are restricted to populations under Hardy–Weinberg equilibrium (HWE; see, e.g., Falconer and MacKay, 1996), including a special case of gene frequency being one half (Mather and Jinks, 1982). Others also adequately account for Hardy–Weinberg disequilibrium (HWD; e.g., Cockerham, 1954; Yang, 2004; Álvarez-Castro and Carlborg, 2007). Regardless of whether a population is in HWE or HWD, Fisher (1918) and others have shown that the additive and dominance genetic effects are simply the coefficient of a linear regression of the genotypic values on the gene content and the deviation from that regression, respectively. The regression coefficient is commonly known as the average effect of substituting one allele by the other in a diploid genotype (Falconer and MacKay, 1996).

As another measure of the additive effect, Fisher (1941) defined the average excess of an allele as the difference by which the average of genotypes carrying that allele exceeds the average of genotypes carrying the alternative allele. Fisher (1941) also pointed out that the average effect is equal to the average excess if the population is in HWE, but it is less than the average excess if inbreeding occurs. Such relationships between average effect and average excess have been subsequently confirmed and elaborated (e.g., Kempthorne, 1957; Falconer, 1985; Templeton, 1987; Lynch and Walsh, 1998).

In this note, we further clarify the relationship between the average effect and the average excess of a gene substitution based on a new set of general contrasts that entail both the average effects and the average excesses as particular cases. We provide a common conceptual and graphical interpretation for both parameters and further dissect how they are related to the decomposition of the genetic variance.

MODEL

Additive and dominance contrasts are commonly used to build and interpret models of genetic effects (e.g., Cockerham, 1954; Li, 1976; Zeng et al., 2005). Such contrasts enter the regression model as:

$$G_{ij} = \mu + \tilde{\alpha} w_{ij} + \tilde{\delta} v_{ij}, \quad (1)$$

where G_{ij} are the genotypic values, μ is the population mean, $\tilde{\alpha}$ and $\tilde{\delta}$ are the additive and dominance genetic effects, and w_{ij} and v_{ij} are, respectively, the coefficients for the additive and dominance contrasts.

In this context, the values 0 and 1 can naturally be used to indicate the presence of alleles A_1 and A_2 in the genotypes, leading to the genotype indicator variable z_{ij} taking the values $z_{11} = 0$, $z_{12} = 1$, and $z_{22} = 2$ for A_1A_1, A_1A_2, and A_2A_2, respectively, and to the coefficients for the additive effects through $w_{ij} = z_{ij} - E(z)$, where $E(z)$ is the expectation of z (see, e.g., Zeng et al., 2005). This indicator variable has thus a clear biological meaning – the gene content of one of the alleles, A_2. When using this indicator variable, the additive parameter is the average effect, i.e., $\tilde{\alpha} = \alpha$, and the dominance parameter is the dominance genetic effect $\tilde{\delta} = \delta$.

On the other hand, the average excesses of alleles in a population under HWD were proffered to further entail the effects of alleles due to correlations with other alleles in that population (Fisher, 1941). Aiming to allow for such correlations in our derivations, we here consider more general indexes. In particular, we introduce a constant c as the ratio of the average effect over the average excess (cf. Eq. 3 of Fisher, 1941). Multiplying z_{ij} by this constant leads to a new genotype indicator variable with $z_{11} = 0$, $z_{12} = c$,

and $z_{22} = 2c$. This new genotype indicator variable will serve to indicate the effective content of allele A_2 in the three genotypes, as it will be further illustrated below.

The use of effective gene contents for obtaining orthogonal contrasts under HWD is summarized in **Table 1**. Obtaining the coefficients for the orthogonal additive contrast, w_{ij}, as $z_{ij} - E(z)$, warrants that $\Sigma p_{ij} w_{ij} = 0$, where p_{ij}, $ij = 11, 12, 22$, are the genotypic frequencies of the population (see, e.g., Cockerham, 1954). The coefficients for the orthogonal dominance contrasts, v_{ij}, are obtained to fulfill $\Sigma p_{ij} v_{ij} = 0$ and $\Sigma p_{ij} w_{ij} v_{ij} = 0$ (Álvarez-Castro and Carlborg, 2007). These are the deviations of the observed genotypic values from the expected values as predicted from the regression of the genotypic values on the effective gene contents.

Additive and dominance contrasts (e.g., the ones built in **Table 1**) can be conveniently expressed in matrix notation. This allows for a straightforward extension of the one locus model to and arbitrary number of loci with arbitrary epistasis under linkage equilibrium (LE; Tiwari and Elston, 1997). It has also been shown that the matrix notation enables straightforward transformations between parameters that have previously been expressed using appropriate contrasts (Álvarez-Castro and Carlborg, 2007).

Let thus **G** be the vector of genetic effects, **E** be the vector entailing the population mean and the additive and dominant parameters and **S** be the genetic-effect design matrix entailing the contrasts that allow for a transformation between vectors **G** and **E**. Then, just using the contrasts in **Table 1** we obtain the matrix expression $\mathbf{G} = \mathbf{S} \cdot \mathbf{E}$ as:

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & -2p_2 c & -\dfrac{p_{12}p_{22}}{2p_1p_2 - 1/2p_{12}} \\ 1 & (p_1 - p_2)c & \dfrac{p_{11}p_{22}}{p_1p_2 - 1/4p_{12}} \\ 1 & 2p_1 c & -\dfrac{p_{11}p_{12}}{2p_1p_2 - 1/2p_{12}} \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \tilde{\alpha} \\ \tilde{\delta} \end{pmatrix}, \quad (2)$$

where p_i, $i = 1, 2$, are the frequencies of the alleles, $p_i = p_{ii} + 1/2p_{ij}$, $j \neq i$.

A UNIFIED FRAMEWORK FOR AVERAGE EFFECTS AND AVERAGE EXCESSES

As mentioned above, the contrasts in **Table 1** provide the average effects of allele substitutions when $c = 1$. It is thus not surprising that in this case Eq. 2 reduces to Álvarez-Castro and Carlborg (2007) Eq. 8 – for the average (additive and dominance) effects. For analyzing how (2) relates to the average excesses, we first recall their definition for one biallelic gene (following Fisher, 1941; Kempthorne, 1957):

$$\begin{cases} \alpha_1^* = \frac{p_{11}}{p_1} G_{11} + \frac{1}{2}\frac{p_{12}}{p_1} G_{12} - \mu \\ \alpha_2^* = \frac{1}{2}\frac{p_{12}}{p_2} G_{12} + \frac{p_{22}}{p_2} G_{22} - \mu \end{cases} \quad (3)$$

By inverting expression (2), it is easy to see that $\tilde{\alpha} = \alpha_2^* - \alpha_1^*$.when $c = 1/(1 + F)$, with $F = 1 - p_{12}/2_{p_1p_2}$ being Wright's (1965) fixation index. F, with the range of $-1 \leq F \leq 1$, reflects any departure from the HWE, toward either an excess or a deficiency of heterozygotes. We can thus rename $\tilde{\alpha} = \alpha^*$, $\tilde{\delta} = \delta^*$ when $c = 1/(1 + F)$. That is to say, Eq. 3 restores the definition of average excesses of the alleles for a biallelic locus. We will consequently refer to (2) with $c = 1/(1 + F)$ as the average-excess formulation of NOIA.

From the general expression (2), we have thus retrieved both the average effects and the average excesses as particular cases of the contrasts in **Table 1**, specifically with $c = 1$ and $c = 1/(1 + F)$, respectively. Therefore, by implementing the effective gene content c we have actually made our model to capture the correlation between alleles that the average excesses account for. Further, using the relationship between the two values of c (1 and $1/(1 + F)$) we are also retrieving the relationship between average effects and average excesses reported by Kempthorne (1957), $\alpha_i = {}^{\alpha_i^*}/_{(1+F)}$, which actually applies to the case of multiple alleles (see also Templeton, 1987).

Evidently, the possible values of the function $1/(1 + F)$ depend on those of the fixation index, F. In particular, $c = 1/(1 + F)$ must always be positive and within the range $1/2 \leq c < \infty$ for the allowable values of F ranging from complete homozygosity ($F = 1$) to complete heterozygosity ($F = -1$). When $F = 0$ (i.e., $c = 1$) we have the well-known case where the average effect and average excess are the same, that is under HWE. Since $c = 1/(1 + F)$ must always be positive, α and α^* will always have the same sign and will verify $|\alpha| = c|\alpha^*|$. Taking all this into account, **Table 2** summarizes how the fixation index affects the relationships between average excesses and additive genetic effects under three situations: heterozygote deficiency ($F < 0$), HWE ($F = 0$) and heterozygote excess ($F > 0$). Within that table, we also stress that the mathematical relationship between average excesses and average effects does not depend upon which one(s) of all potential biological features is (are) underlying a particular set of observed genotype frequencies.

Table 1 | Coefficients of orthogonal contrasts for the average effects and the average excesses for two allels at a locus.

Genotypes	Frequencies	z_{ij}	$w_{ij} = z_{ij} - E(z)$	v_{ij}
A_1A_1	p_{11}	0	$-2\ p_2\ c$	$-\dfrac{p_{12}p_{22}}{2p_1p_2 - 1/2p_{12}}$
A_1A_2	p_{12}	c	$(p_1 - p_2)\ c$	$\dfrac{p_{11}p_{22}}{2p_1p_2 - 1/4p_{12}}$
A_2A_2	p_{22}	$2c$	$2\ p_1 c$	$-\dfrac{p_{11}p_{12}}{2p_1p_2 - 1/2p_{12}}$

The non-zero constant c is introduced for accounting for effective gene contents.

Table 2 | Summary of some relevant mathematical and biological features associated to different statuses of the heterozygosity of a population.

Heterozygotes deficiency	Observed heterozygotes fit HWE	Heterozygotes excess
$0 < F \leq 1$	$F = 0$	$-1 \leq F < 0$
$1/2 \leq c < 1$	$c = 1$	$c > 1$
$\|\alpha^*\| > \|\alpha\|$	$\alpha^* = \alpha$	$\|\alpha^*\| < \|\alpha\|$
Assortative mating or homozygotes favored or population structure	Random mating and either no selection or geometric fitnesses	Dissassortative mating or heterozygotes favored or gene duplication

PARTITIONING THE GENOTYPIC VALUES AND THE GENETIC VARIANCE

The average-excess formulation [expression (2) with $c = 1/(1 + F)$] comes from a linear regression (1) and it can thus be expressed by means of its intercept, μ, and its regression coefficient, α^*, as:

$$\hat{G}(w) = \mu + \alpha^* w \tag{4}$$

This regression entails a decomposition of the genotypic values in which the predictions from the regression are the additive components and the deviations of the regression – due to dominance interactions – are the dominance components. For instance, the predicted [by (4)] value for genotype A_1A_1 is $\alpha^*_{11} = \hat{G}(-c)$. Now, both **Table 1** and expression (2) show that the dominance contrasts, v_{ij}, do not depend upon the scaling factor c and, hence, they are equal for the statistical and the statistical excess formulations. This implies that the dominance deviations are the same in both cases, i.e., $\delta^*_{ij} = \delta_{ij}$ and that, therefore, $\alpha^*_{ij} = \alpha_{ij} = \alpha_i + \alpha_j$. That is to say, both formulations lead to the same decomposition of genotypic values,

$$G_{ij} = \mu + \alpha_{ij} + \delta_{ij} \equiv \mu + \alpha^*_{ij} + \delta^*_{ij}. \tag{5}$$

This is illustrated in **Figure 1**, where we show the graphical interpretation of the decomposition of genotypic values coming from the average excesses and compare it with the classical decomposition coming from the average effects (Fisher, 1918). Note, particularly, that the decomposition of genotype A_1A_1 into additive and dominance parts is the same regardless of which linear regression is used. Interestingly, although for the average effects formulation (with $c = 1$) the predictions of the regression can be obtained by just summing up the appropriate average effects (see, e.g., Álvarez-Castro and Carlborg, 2007), this does not hold for the average-excess formulation [with $c = 1/(1 + F)$], i.e., $\alpha^*_{ij} \neq \alpha^*_i + \alpha^*_j$, unless the genotypic frequencies are under HWE. The reason for this is also noted in **Figure 1**, where it can be seen that α^*_i and α^*_{ij} associated to different values for the regression independent variable ($\alpha^*_{11} = \hat{G}(-c)$ whereas $\alpha^*_1 + \alpha^*_1 = \hat{G}(-1)$). The exact relationship between these values under HWD is straightforward from $\alpha^*_{ij} = \alpha_{ij} = \alpha_i + \alpha_j$ and $\alpha_i = c\alpha^*_i$ (Kempthorne, 1957), which lead to:

$$\alpha^*_{ij} = c\left(\alpha^*_i + \alpha^*_j\right). \tag{6}$$

The decomposition of genotypic values being the same for the average effects and the average excesses (5) necessarily implies that they also lead to the same decomposition of the genetic variance. We have confirmed this result by substituting the average-excess additive contrasts (**Table 1**, with $c = 1/(1 + F)$) in the equation for the additive variance (see, e.g., Cockerham, 1954). When doing so, a common factor c^2 can be simplified from both the numerator and the denominator of that expression so that the original expression for the additive variance is retrieved.

The additive variance coming from the average excesses is the variance of the values α^*_{ij}. Thus, the average excesses of the alleles enter the computation of the additive variance by just applying

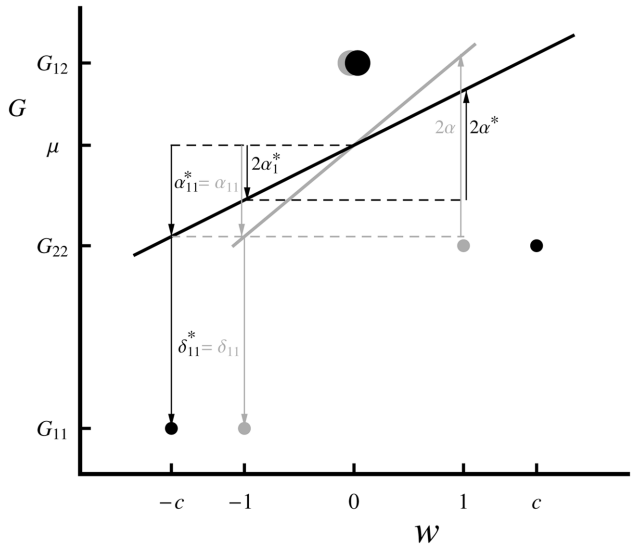


FIGURE 1 | Graphical interpretation of the decomposition of the genotypic values (5) through the statistical excess (in black) and the statistical (in gray) formulations of NOIA for one locus with two alleles. For simplicity, a case with equal allele frequencies ($p_1 = p_2 = 1/2$) is shown. The specific genotypic values (circles; $G_{11} = 1$, $G_{12} = 3$, $G_{22} = 2$) displaying overdominance and a fixation index ($F = -2/5$) have been chosen for facilitating the visualization of the parameters of interest. The size of the circles represents the frequency of the genotypes. Horizontal dashed lines emphasize coincident arrow edges, the upper one corresponding to the population mean phenotype, $\mu = 2.55$. The regression independent variable of the statistical formulation is the gene content, whereas the one of the statistical excess formulation is scaled by $c = 1/(1 + F) = 5/3$ and it works as an effective gene content. For both cases, the independent variable, w, is rescaled by its expectation as shown in **Table 1**.

(6). Although a common way to express and compute the additive variance under HWD entails both the average (additive) effects and the average excesses [see, e.g., expression (4.23a) in Lynch and Walsh, 1998], here we have shown that either formulation alone suffices to provide the additive variance under HWD. We recall that this is true as long as the formulations are built using contrasts that are appropriate to HWD – as the ones we are providing in this communication for both the biallelic case.

EFFECTIVE GENE CONTENT

Hardy–Weinberg disequilibrium implies that alleles become (either positively or negatively) correlated in zygotes as compared to the expected genotype frequencies under HWE. A deficiency of heterozygotes, for instance, causes alleles to become positively correlated, leading to their effective additive contribution to the genotypes of a population to be more extreme (i.e., further away from their expectation) than under HWE. Fisher (1941) noted that this is accounted for by the average excesses. We note that this is not in contradiction with the interpretation of the average excesses of one allele as the conditional average genotypic deviation of the individuals that received that allele from at least one parent (see, e.g., Templeton, 2006).

For the biallelic case, we can trace Fisher's (1941) remark in our graphical interpretation (**Figure 1**). We first recall that although

both the average effects and the average excesses are linear regressions of the genotypic values (the regression dependent variable) as expressed in (1), each of them is regressed on a different independent variable. The independent variable of the formulation of average effects is the actual content of allele A_2 (which is in **Figure 1** shown as rescaled by its expectation) whereas the independent variable of the average-excess formulation is the effective content of allele A_2 measured by a factor c. This factor being greater than one in our example ($c = 5/3$) reflects an excess of heterozygotes (particularly with $F = -2/5$) and makes the slope of the regression for the average excess, α^*, to be less steep than the one on the actual gene content, α, as noted in **Table 2**. Conversely, a deficiency of heterozygotes would make the slope of the average-excess regression to become steeper than the one of the regression for the average effects. Thus, the effective gene content c leads to the average excesses to reflect the effective contributions of the alleles to the genotypes of a population.

CLOSING PERSPECTIVE

In conclusion, we have showed here that Fisher's (1941) definition of average excesses can be phrased within a new regression framework that also generalizes the average effects. This has enabled us to clarify the significance of the average excesses in different ways. First, we have expressed the average excesses in terms of matrix notation within the NOIA framework, which entails the extension of that theory to multiple loci with arbitrary epistasis under LE and allows us to easily transform between average excesses and other genetic parameters. Second, we have fully integrated the average excesses into the theory for the decomposition of the genotypic values and the genetic variance into additive and dominant components. Third, we have provided a graphical interpretation of the average excesses that is analogous to the one of the average effects. Finally, we interpret the factor determining the relationship between average effects and average excesses as the effective gene content of individuals, accounting not only for the effects of their alleles but also for how pairs of alleles are correlated in a particular population.

ACKNOWLEDGMENTS

José M. Álvarez-Castro acknowledges funding by an "Isidro Parga Pondal" contract from the autonomous administration Xunta de Galicia. This research has been supported by project BFU2010-20003 form the Spanish Ministry of Science (José M. Álvarez-Castro) and the Natural Sciences and Engineering Research Council of Canada, Grant OGP0183983 (Rong-Cai Yang).

REFERENCES

Álvarez-Castro, J. M., and Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176, 1151–1167.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39, 859–882.

Falconer, D. S. (1985). A note on Fisher's 'average effect' and 'average excess'. *Genet. Res. (Camb.)* 46, 337–347.

Falconer, D. S., and MacKay, T. F. C. (1996) *Introduction to Quantitative Genetics*, 4th Edn. Harlow: Prentice Hall.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* 52, 339–433.

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Ann. Eugen.* 11, 53–63.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. New York: Wiley.

Li, C. C. (1976). *First Course in Population Genetics*. Pacific Grove, CA: The Boxwood Press.

Lynch, M., and Walsh, B. (1998). *Genetic Analysis of Quantitative Traits*. Sunderland: Sinauer.

Mather, K., and Jinks, J. L. (1982). *Introduction to Biometrical Genetics*. London: Chapman and Hall.

Templeton, A. R. (1987). The general relationship between average effect and average excess. *Genet. Res. (Camb.)* 49, 69–70.

Templeton, A. R. (2006). *Population Genetics and Microevolutionary Theory*. Hoboken, NJ: John Wiley & Sons.

Tiwari, H. K., and Elston, R. C. (1997). Deriving components of genetic variance for multilocus models. *Genet. Epidemiol.* 14, 1131–1136.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19, 395–420.

Yang, R.-C. (2004). Epistasis of quantitative trait loci under different gene action models. *Genetics* 167, 1493–1505.

Zeng, Z. B., Wang, T., and Zou, W. (2005). Modeling quantitative trait Loci and interpretation of models. *Genetics* 169, 1711–1725.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 December 2011; accepted: 17 February 2012; published online: 09 March 2012.

Citation: Álvarez-Castro JM and Yang R-C (2012) Clarifying the relationship between average excesses and average effects of allele substitutions. Front. Gene. ***3****:30. doi: 10.3389/fgene.2012.00030*

This article was submitted to Frontiers in Genetic Architecture, a specialty of Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 22 July 2014
doi: 10.3389/fgene.2014.00227

Analysis of linear and non-linear genotype × environment interaction

Rong-Cai Yang[1,2]*

[1] Alberta Agriculture and Rural Development, Edmonton, AB, Canada
[2] Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Keyan Zhao, University of California, Los Angeles, USA
Urko M. Marigorta, Georgia Institute of Technology, USA

***Correspondence:**
Rong-Cai Yang, Department of Agricultural, Food and Nutritional Science, University of Alberta, 410 Agriculture/Forestry Centre, Edmonton, Alberta T6G 2P5, Canada
e-mail: rong-cai.yang@ales.ualberta.ca

The usual analysis of genotype × environment interaction (G × E) is based on the linear regression of genotypic performance on environmental changes (e.g., classic stability analysis). This linear model may often lead to lumping together of the non-linear responses to the whole range of environmental changes from suboptimal and super optimal conditions, thereby lowering the power of detecting G × E variation. On the other hand, the G × E is present when the magnitude of the genetic effect differs across the range of environmental conditions regardless of whether the response to environmental changes is linear or non-linear. The objectives of this study are: (i) explore the use of four commonly used non-linear functions (logistic, parabola, normal and Cauchy functions) for modeling non-linear genotypic responses to environmental changes and (ii) to investigate the difference in the magnitude of estimated genetic effects under different environmental conditions. The use of non-linear functions was illustrated through the analysis of one data set taken from barley cultivar trials in Alberta, Canada (Data A) and the examination of change in effect sizes is through the analysis another data set taken from the North America Barley Genome Mapping Project (Data B). The analysis of Data A showed that the Cauchy function captured an average of >40% of total G × E variation whereas the logistic function captured less G × E variation than the linear function. The analysis of Data B showed that genotypic responses were largely linear and that strong QTL × environment interaction existed as the positions, sizes and directions of QTL detected differed in poor vs. good environments. We conclude that (i) the non-linear functions should be considered when analyzing multi-environmental trials with a wide range of environmental variation and (ii) QTL × environment interaction can arise from the difference in effect sizes across environments.

Keywords: barley, environmental index, estimation, genotype × environment interaction, non-linear functions, quantitative trait loci

INTRODUCTION

Inconsistent performance of genotypes over different environments known as genotype × environment interaction (G × E) remains to be a major impediment to genetic improvement of biological species in Canada and elsewhere. G × E is particularly important for plant species (e.g., agricultural crops and forest trees) because they spend their entire life at the same locality. Over the past decades, the assessment of G × E has been done with the data obtained from testing of the same genotypes over multiple environments (locations or years), i.e., multi-environmental trials (Yang, 2007).

The G × E effect has been incorporated into quantitative genetic models (Falconer and Mackay, 1996) through the use of genetic correlations within and between individual genotypes (e.g., Crossa et al., 2004; Burgueño et al., 2008). The basic idea behind such an approach is to predict genetic values through borrowing information among individuals from genetic relationships, and within individuals (across environments) from genetic and environmental correlations. The analysis of such correlation structure has been performed to obtain the parsimony description of G × E variation using different versions of linear-bilinear models based on a mathematical technique known as singular value decomposition (SVD) (Golub and Reinsch, 1970). One popular use of the SVD technique is the biplot analysis of G × E based on the two commonly used rank-two linear-bilinear models: the additive main effects and multiplicative interaction (AMMI) model and the genotype main effects and genotype × environment interaction effects (GGE) model (i.e., fitted to residuals after removal of environment main effects) (for review, see Yang et al., 2009). Recently, Burgueño et al. (2008) and Cullis et al. (2010) described a similar biplot analysis under a mixed-model framework using a series of rank-two factor-analytic (FA) model. Apart from the adequacy of the rank-two models and other statistical issues, Yang et al. (2009) pointed out that the biplot analysis has contributed little to our understanding of the nature of G × E variation because it is a descriptive analysis with little predictive power.

Baker (1988) and others (e.g., Scheiner, 1993; Lindgren and Ying, 2000) have suggested the use of predictive models based on linear and non-linear response functions for studying G × E. The classic stability analysis based on simple linear regression model as pioneered by Yates and Cochran (1938) is a special case of the general non-linear predictive models. In addition, linear functions would usually account for a small portion of G × E variation if a wide range of environmental conditions are tested. On the other hand, for quantitative traits such as crop yield or human complex diseases (Franks et al., 2013), the G × E is manifested when the magnitude of the genetic effect differs across the range of environmental conditions regardless of whether the response to environmental changes is linear or non-linear. For this reason, many recent genome-wide association studies (GWAS) in human (Kilpelainen et al., 2011; Qi et al., 2012) have focused on determining the effect sizes of causal variants (e.g., SNPs) over different environmental conditions (e.g., different lifestyle behaviors).

The objectives of this paper are two folds. First, we investigate the use of different non-linear functions for modeling genotypic response to environmental changes or gradients. In this case, G × E is present when the response curves fail to be parallel (Baker, 1988). Similar concept has been used in evolution and ecology but under different names [e.g., phenotypic plasticity (robustness), reaction norm] (e.g., Via et al., 1995). Second, we examine whether there are differences in estimated genetic effects under different environmental conditions. It is generally expected that a larger effect is more likely found in the environmental condition where the expression of a gene is facilitated than in the environmental condition where the expression of a gene is not facilitated.

MATERIALS AND METHODS

DESCRIPTION OF NON-LINEAR FUNCTIONS

As a starting point, we provide a brief description of the classic stability analysis that is based on a linear regression function (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966; Perkins and Jinks, 1968):

$$y_{ij} = a_i + b_i x_j \tag{1}$$

Where y_{ij} is the performance (say yield) of the ith genotype tested in jth environment, x_j is the mean yield of all genotypes tested in the jth environment (known as environmental index), the intercept a_i is the yield of the ith genotype at the worst environment, and the slope b_i measures the stability of the ith genotype.

According to Finlay and Wilkinson (1963), all genotypes can be conveniently classified into three groups: (i) genotypes with average stability ($b_i = 1.0$); (ii) genotypes with low stability or high sensitivity to environmental changes ($b_i > 1.0$) and (iii) genotypes with high stability or low sensitivity to environmental changes ($b_i < 1.0$). Eberhart and Russell (1966) further refined this definition by suggesting that a stable genotype would be the one with average stability, low variance due to deviations from regression and high mean yield.

However, linear response usually accounts for only a small portion of the G × E variation and the responses are most often non-linear in practice (Knight, 1973; Jinks and Pooni, 1988). This occurs because when individuals of the same genotype are evaluated at different levels of an environmental factor ranging from suboptimal, optimal to super-optimal levels, their performance (i.e., yield) often shows a continuous non-linear relationship with the environment. The response curve can rise from near zero performance at extreme suboptimal levels of the environmental factor to some asymptotic value at optimal levels, and then decrease to near zero value at extreme super-optimal levels. If a small portion of the environmental range is evaluated, only the linear response could possibly be observed within this limited range of environmental conditions.

Here we briefly describe some well-known non-linear functions that have been used to model relationships of yield or growth with a single more defined environmental variable (for details, see Baker, 1988; Ratkowsky, 1993). The most obvious non-linear function is a quadratic function (parabola function) and it is often used to describe the relationship between grain yield and field water availability (e.g., McKenzie et al., 2004):

$$y_{ij} = a_i + b_i x_j + c_i x_j^2 \tag{2}$$

The quadratic function has been also used to describe the genetic response to climate variables in forest trees (Rehfeldt et al., 1999). Another non-linear function is the reciprocal of the quadratic function used to describe the relationship between yield and planting density (Baker, 1988):

$$y_{ij}^{-1} = a_i + b_i x_j + c_i x_j^2 \tag{3}$$

This general expression can take several special forms, one of which is known as Cauchy function,

$$y_{ij} = \frac{k_i}{\left[1 + \frac{(x_j - x_{\max})^2}{r_i^2}\right]} \tag{4}$$

Where K_i is a parameter that scales yield from zero to one (i.e., $0 \leq K_i \leq 1$), $x_{\max}$ is the x value at which the maximum yield is achieved and γ_i is the scale parameter which measures the range of genotypic response to environmental changes. This Cauchy function has been used to delineate breeding zones in forest trees (Raymond and Lindgren, 1990; Lindgren and Ying, 2000). The logistic curve:

$$y_{ij}^{-1} = a_i + b_i c_j^{x_j} \tag{5}$$

is often used to describe the plant growth with age, but it can also be useful for the response to the environmental changes (Baker, 1988; West et al., 2001; Zuo et al., 2012). Roberds and Namkoong (1989) proposed the use of the Gaussian function to model the genotypic response to an environmental gradient:

$$y_{ij} = \frac{k_i}{\sqrt{2\pi r_i^2}} e^{\left[\frac{(x_j - x_{\max})^2}{2r_i^2}\right]} \tag{6}$$

When $K_i = 1$, Equation (6) becomes the normal probability density function. These non-linear functions are graphed in **Figure 1**.

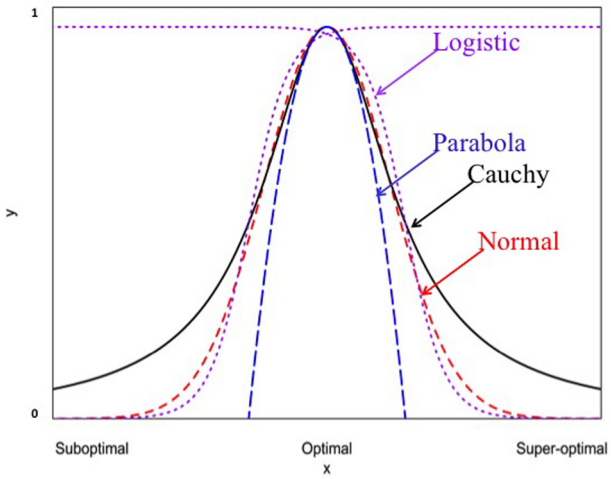


FIGURE 1 | Four different non-linear functions for studying genotype-environment interaction (normal, Cauchy, parabola, and logistic).

It should be noted that the y-axis and x-axis in **Figure 1** are rescaled in standardized units. For example, the standardized Cauchy function is given by:

$$y'_{ij} = \frac{1}{1 + x'^2_{ij}} \tag{7}$$

Where $y'_{ij} = \frac{y_{ij}}{k_i}$ and $x'_{ij} = \frac{x_j - x_{\max}}{r_i}$ Thus, y'_{ij} becomes a relative measure of the performance within the range of 0 (0%)–1 (100%). All non-linear functions are indistinguishable at or near the optimum $x'_{ij} = 0$. For example, the Cauchy function can be well approximated by a quadratic function at the rescaled axises because of the following mathematical relationship:

$$\frac{1}{1 + x'^2_{ij}} \to 1 - x'^2_{ij} \text{when} x'_{ij} \to 0 \tag{8}$$

but the approximation becomes less desirable at the extreme environmental conditions (i.e., $|x'_{ij}| >> 0$).

ANALYSIS OF EMPIRICAL DATA

We will describe the analysis of two empirical data sets. The first data set (Data A) is taken from Yang et al. (2006) who analyzed 324 replicated barley cultivar trials sown at 84 sites across three provinces (Alberta, Saskatchewan and Manitoba) in the Canadian prairies during 1995–2003. Here we illustrate the use of non-linear G × E analysis of the data taken from the trials in the province of Alberta only. The data set for the analysis is briefly recapitulated now. In each year, there were 16 (1995)–22 (2000) trials planted at different locations across Alberta. Each trial consisted of 39–44 barley cultivars. It should be pointed that in a given year, the same cultivars were usually included in each and every trial but over different years, at least some cultivars were different in the same and different test sites either due to a turnover to newly registered cultivars or to unavailability of pedigree seed of older cultivars. The same check cultivars were used across the different years. All trials were conducted using a randomized complete block design with three or four replications. Cultural practices such as fertility, tillage and pest control varied from site to site but were considered to be the most appropriate for the individual sites.

Following the procedure of Yang et al. (2006), the usual analysis of variance partitioned the total sum of squares in each year into components due to the site effects (E), the cultivar effects (G) and the interaction between cultivar and site effects (G × E) using SAS PROC MIXED (Sas Institute Inc, 2012). Further partitioning of the G × E variation under different non-linear functions was carried out using appropriate data transformations that enabled the analysis of non-linear G × E under the mixed-model framework. The different non-linear functions were compared interms of their ability to capture the amount of G × E variation.

The second data set (Data B) is a publicly available data set that we previously analyzed using single-marker analysis (Ham et al., 2010) and genome-wide prediction (Yang and Ham, 2012). The data set consisted of 150 doubled haploid (DH) lines that were developed from a cross between two malting barley varieties (Steptoe × Morex) for the North American Barley Genome Mapping Project (NABGMP) (http://wheat.pw.usda.gov). These DH lines were tested in 16 environments over North America for yield and seven other agronomic and malt quality traits. A total of 223 restricted fragment length polymorphism (RFLP) makers mapped over the seven chromosomes of the barley genome with 37, 37, 31, 33, 29, 22, and 34 makers being mapped on chromosomes 1, 2, 3, 4, 5, 6, and 7, respectively. The effects of these RFLP markers were estimated using a R package, GLMNET/R, at three representative environments: poor (minimum environmental index), average (mean environmental index) and good (maximum environmental index) environments. GLMNET/R implemented an efficient procedure for fitting the entire elastic-net regularization path for super-saturated linear regression as in genome-wide association studies (GWAS) (Friedman et al., 2010; R Core Team, 2012). The elastic-net penalty (P_α) is a compromise between the ridge-regression penalty ($\alpha = 0$) and the LASSO penalty ($\alpha = 1$), where α is related to the degree of shrinkage of marker effects. Two shrinkage methods, elastic net with $\alpha = 0.5$ and $\alpha = 1$ (i.e., LASSO), were used for genome-wide estimation of marker effects on response at poor, average and good environments.

RESULTS

DATA A

We (Yang et al., 2006) previously partitioned the total variability into components due to genotypes (G), environments (E) and G × E, and G × E accounted for 6.6% (2003)–23.9% (2000) of the total variability across different years. Here we further partitioned the G × E variability into a component that could be explained by different linear and non-linear models described above and a residual (**Table 1**). This further partitioning was based on linear or non-linear regression of yield on the environmental index (calculated as the mean of all cultivars at each and every test location). It is evident from **Table 1** that different non-linear models

Table 1 | Percentages of genotype × environment interaction variation explained by linear function and four non-linear functions in barley cultivar trials in Alberta tested in 1995–2003.

Year	Linear	Logistic	Parabola	Normal	Cauchy
1995	8.49	7.52	11.10	11.47	20.17
1996	8.84	7.32	14.14	13.06	25.28
1997	6.72	5.88	11.81	9.97	12.54
1998	8.40	7.70	13.15	15.12	26.54
1999	14.70	15.75	20.41	20.85	36.56
2000	5.91	8.34	8.67	14.30	32.39
2001	6.95	11.77	13.16	35.04	86.45
2002	23.60	13.17	40.08	33.46	84.87
2003	17.71	14.06	22.51	18.88	37.69
Average	11.26	10.17	17.23	19.13	40.28

captured different amounts of the total G × E variation, ranging from an average of 10.2% for logistic model to 40.3% for Cauchy model. It is somewhat surprising that some non-linear models (e.g., logistic model) actually captured less G × E variation than the linear model. For a given model, there was also a large amount of year-to-year variation in the percentages of the G × E variation being captured. For example, Cauchy model captured 12.5% in 1997 and 86.5% in 2001. This result suggests that G × E variation is more predictable in some "good" years than in other "poor" years. In good years, stable and non-extreme weather or other agroclimatic conditions are available for optimal performance of individual genotypes whereas in poor years, such conditions do not exist.

DATA B

Responses of the DH lines to environmental index were examined under different linear and non-linear models. The responses of most DH lines were linear (**Figure 2**). Furthermore, the variation in such linear response was greater in "good" environments (i.e., the locations with higher environmental index values) than in "poor" environments (i.e., the locations with lower environmental index values). It is evident from **Figures 3**, **4** that Elastic net ($\alpha = 0.5$) detected more marker effects than LASSO ($\alpha = 1.0$) but LASSO gave much sharper resolution of marker effects. Under both estimation methods, marker effects were more pronounced in good environment than in poor environment.

DISCUSSION

Differential responses of genotypes to environmental conditions (G × E interactions) can be linear or non-linear. Most current analyses of such responses are limited to the use of linear models. In this study, we explore the use of different non-linear models for characterizing and dissecting G × E interaction. This was done by extending the linear regression on environmental indexes (the means of all genotypic values at individual environments) or the classic stability analysis (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966; Perkins and Jinks, 1968) to the non-linear regression analysis. In the past, several non-linear functions including logistic,

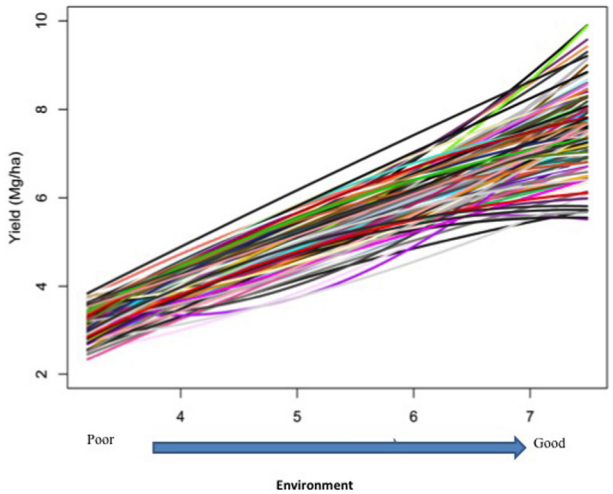


FIGURE 2 | Responses of 150 doubled-haploid lines of barley from a cross between two malting barley cultivars (Steptoe × Morex) for the North American Barley Genome Mapping Project (NABGMP). The range of the environmental index values runs from low (poor environment) to high (good environment).

quadratic (parabola), Cauchy and normal functions have been individually used to describe genotypic responses to environments (e.g., Knight, 1973; Jinks and Pooni, 1979; Roberds and Namkoong, 1989; Raymond and Lindgren, 1990; Van Tienderen and Koelewijn, 1994; Lindgren and Ying, 2000). For example, Van Tienderen and Koelewijn (1994) found that the quadratic function was "statistically significantly better" than the linear function. In this study, our comparison of these representative non-linear functions (**Figure 1**) reveals the following characteristics. First of all, when the parameters are appropriately chosen or rescaled, the response curves of different non-linear functions near the optimum are indistinguishably similar, but their differences become increasingly evident when the environmental condition is not good (suboptimal) or too good (super-optimal). Second, should the true response be non-linear but be treated as linear, it would be difficult to tell the difference between non-linear responses to suboptimal and super-optimal conditions because in the linear analysis, both suboptimal and super-optimal conditions are lumped together to represent a deteriorated environment (**Figure 5**). Thus, the linear analysis would cause the reduced range of environmental variation when non-linear response is present but its presence unknown to the researcher or simply ignored! Third, including responses to both suboptimal and super-optimal conditions provides more opportunities to characterize the nature of G × E interaction. For example, differences in the rate of increase in response at suboptimal levels would reflect differences in efficiency but differences in the rate of decrease in response at super-optimal levels would reflect differences in tolerance.

It may not totally surprising from this study that the Cauchy function is the best in capturing the G × E variation because it may be best representative of how different genotype respond to the whole range of environmental conditions. Each genotype

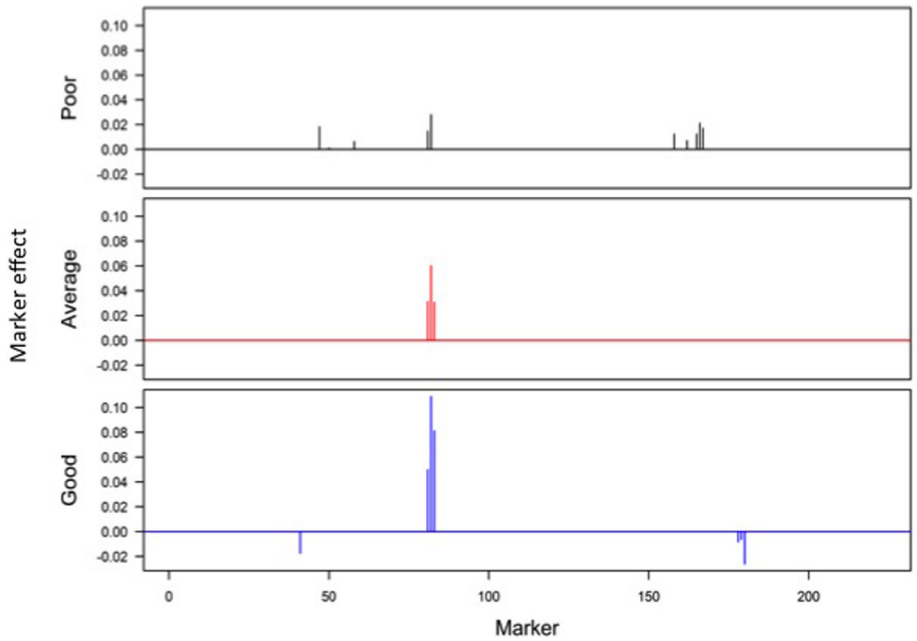


FIGURE 3 | Genome-wide scan of QTLs responsible for barley yield in poor, average, and good environments using the ridge regression analysis.

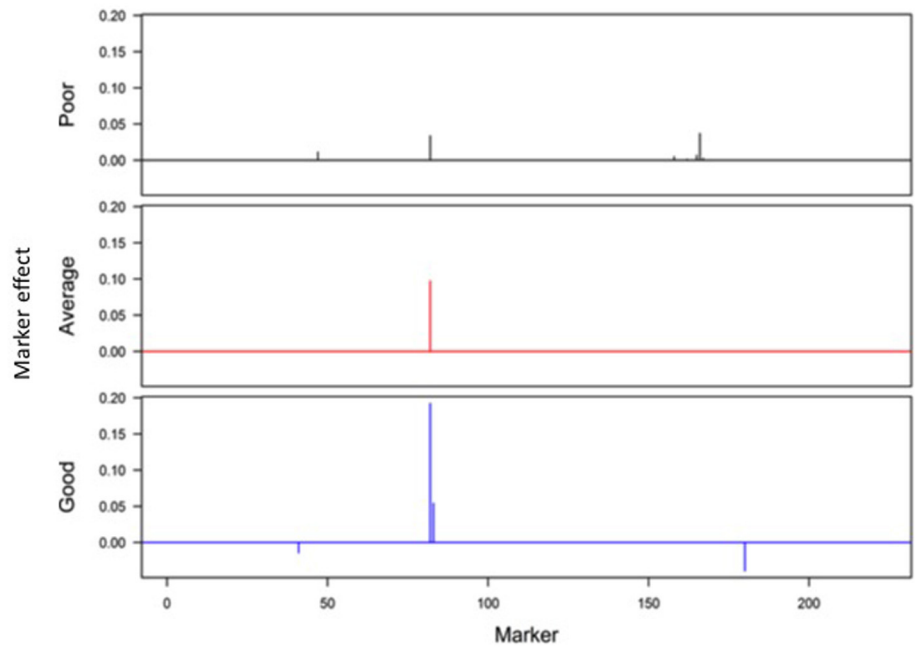


FIGURE 4 | Genome-wide scan of QTLs responsible for barley yield in poor, average, and good environments using the LASSO analysis.

would have its own optimal growing environment. Any deviation from such optimum, either super-optimal or sub-optimal conditions, would cause a reduced performance or adaptation. The reduction must be very gentle for relatively mild super-optimal or sub-optimal conditions. For the extremely poor environments, the reduction asymptotically approaches a nonzero minimum. This scenario is best described by the Cauchy function which has a gentle decline at the regions close to the optimum (the center) and it has very long, flat tails at either side of the center but never converges. Comparing to the other non-linear functions, the Cauchy function is more sensitive to the values close to the optimum but less sensitive to the values at extreme environments which are of

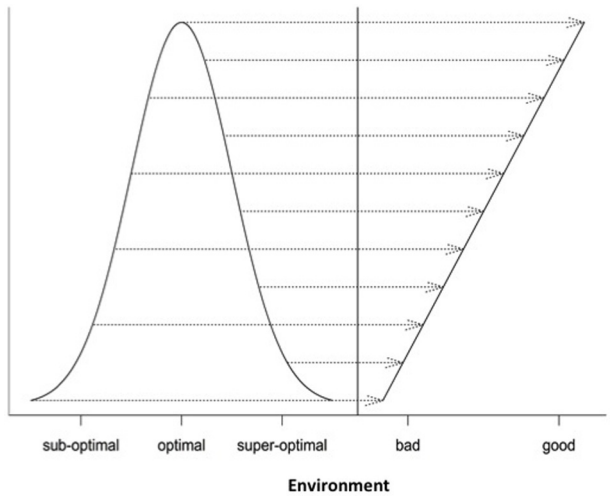


FIGURE 5 | A demonstration of masking true (non-linear) responses to environmental changes if a linear function is used.

little practical interest (Raymond and Lindgren, 1990; Lindgren and Ying, 2000). Thus the Cauchy should be considered in future plant and animal breeding and evolution studies.

Our analysis of Data A shows that different non-linear functions captured different amounts of G × E interaction variation with Cauchy function capturing an average of 40% of the total G × E variation which is twice the amount captured by the second best model (normal function). This striking capability of Cauchy function was also observed in Raymond and Lindgren (1990) and Lindgren and Ying (2000). It is evident from **Figure 1** that all non-linear functions are similar and indistinguishable when environmental conditions are close to the optimum but they become markedly different when environmental conditions move toward the extremes. Our results suggest that the actual range of environmental conditions as represented by all test locations over the years is too extended to be accommodated by all the functions except for the Cauchy function which can accommodate the environmental conditions at some distance away from the optimum. Thus, in practical applications, the choice of a non-linear function should be done after examining the actual distributions of environmental conditions either from previous experiences or from empirical data. It should also be reminded that a sufficient number of environments (e.g., ∼40 locations in our study) are needed so that the true distribution of environmental conditions can be well approximated by the empirical data.

The results from the analysis of Data B reveal that responses of 150 DH lines to environmental indexes were largely linear (**Figure 2**). The 16 environments (essentially 12 locations in 2 years) at which these DH lines were tested would hardly be considered sufficient for covering the whole environmental range. Thus, the linear responses may be reflective of the response to a limited range of environmental indexes. The possibility of non-linear responses could not be ruled out particularly if the whole environmental range is available. Even within this limited environmental range, our analysis revealed some interconnected and interesting features. First of all, the variation in the responses of DH lines was greater in good environment than in poor environment. Second, the contrast between good and poor environments correspondingly led to the difference in the estimated positions, sizes and directions of QTL effects between these environments and this occurred irrespective of which method was used (**Figures 3**, **4**). Third, inconsistency in the positions, sizes and directions of QTLs across the environmental range is a direct evidence of strong QTL × environment interaction.

As just mentioned above, there is increase in the effect size of detected QTLs in good environment in comparison to poor environment (**Figures 3**, **4**). Similar observations have recently been made in many human GWAS particularly with respect to GWAS-discovered causal SNPs controlling the susceptibility of obesity. For example, Kilpelainen et al. (2011) showed that the risk effect of FTO (fat mass and obesity associated) alleles was about 100% and larger in physically inactive individuals than in active individuals from North America. Similar increase in the effect size was observed when individuals with ≥1 serving sugar-sweetened beverage per day were compared to those with sugary beverage intake <1 serving per month (Qi et al., 2012). Such increase in the effect size occurs because there are causal variants that lead to more phenotypic variation in the inactive lifestyle than in the active lifestyle. While generally being ignored in the past, our study and those other recent studies raise an important point that the genetic effects must not only be defined and estimated under a reference population, but also under an appropriate environment.

In conclusion, this paper calls for the attention to the use of non-linear functions for studying G × E interaction. We illustrate that the portion of G × E variation due to non-linear responses can be substantial if the correct non-linear function is used. We also emphasize that the correct identification of non-linear functions depends critically on how close the estimated environmental range is to the true range.

ACKNOWLEDGMENTS

I thank Dr. Zhiqiu Hu for computational and technical assistance, and two anonymous reviewers for helpful comments. This research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC OGP0183983).

REFERENCES

Baker, R. J. (1988). "Differential response to environmental stress," in *Proceedings of the Second International Conference on Quantitative Genetics*, eds B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong (Sunderland, MA: Sinauer Associates), 492–504.

Burgueño, J., Crossa, J., Cornelius, P. L., and Yang, R.-C. (2008). Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. *Crop Sci.* 48, 1291–1305. doi: 10.2135/cropsci2007.11.0632

Crossa, J., Yang, R.-C., and Cornelius, P. L. (2004). Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. *J. Agric. Biol. Environ. Stat.* 9, 362–380. doi: 10.1198/108571104X4423

Cullis, B. R., Smith, A. B., Beeck, C. P., and Cowling, W. A. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II. exploring variety by environment interaction using factor analysis. *Genome* 53, 1002–1016. doi: 10.1139/G10-080

Eberhart, S. T., and Russell, W. (1966). Stability parameters for comparing varieties. *Crop Sci.* 6, 36–40. doi: 10.2135/cropsci1966.0011183X000600010011x

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics.* New York, NY: Longman.

Finlay, K., and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Aust. J. Agric. Res.* 14, 742–754. doi: 10.1071/AR9630742

Franks, P. W., Pearson, E., and Florez, J. C. (2013). Gene-environment and gene-treatment interactions in type 2 diabetes. *Diabetes Care* 36, 1413–1421. doi: 10.2337/Dc12-2211

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1.

Golub, G. H., and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numer. Math.* 14, 403–420. doi: 10.1007/BF02163027

Ham, B. J., Spaner, D., Rahman, M. H., Yeh, F. C., and Yang, R.-C. (2010). Analysis of genotype-environment interactions from a genome-wide survey of quantitative trait loci in a barley population. *Curr. Top. Genet.* 4, 21–32.

Jinks, J. L., and Pooni, H. S. (1979). Non-linear genotype × environment interactions arising from response thresholds. *Heredity (Edinb.)* 43, 57–70. doi: 10.1038/hdy.1979.59

Jinks, J., and Pooni, H. (1988). "The genetic basis of environmental sensitivity," in *Proceedings of the Second International Conference on Quantitative Genetics*, eds B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong (Sunderland, MA: Sinauer), 505–522.

Kilpelainen, T. O., Qi, L., Brage, S., Sharp, S. J., Sonestedt, E., Demerath, E., et al. (2011). Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS Med.* 8:e1001116. doi: 10.1371/journal.pmed.1001116

Knight, R. (1973). The relation between hybrid vigour and genotype-environment interactions. *Theor. Appl. Genet.* 43, 311–318. doi: 10.1007/BF00275258

Lindgren, D., and Ying, C. (2000). A model integrating seed source adaptation and seed use. *New Forests* 20, 87–104. doi: 10.1007/BF00275258

McKenzie, R. H., Middleton, A. B., Hall, L., Demulder, J., and Bremer, E. (2004). Fertilizer response of barley grain in south and central Alberta. *Can. J. Soil Sci.* 84, 513–523. doi: 10.4141/s04-013

Perkins, J. M., and Jinks, J. (1968). Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. *Heredity (Edinb.)* 23, 339. doi: 10.1038/hdy.1968.48

Qi, Q. B., Chu, A. Y., Kang, J. H., Jensen, M. K., Curhan, G. C., Pasquale, L. R., et al. (2012). Sugar-sweetened beverages and genetic risk of obesity. *N. Engl. J. Med.* 367, 1387–1396. doi: 10.1056/NEJMoa1203039

Ratkowsky, D. A. (1993). Principles of nonlinear regression modeling. *J. Ind. Microbiol.* 12, 195–199. doi: 10.1007/BF01584190

Raymond, C. A., and Lindgren, D. (1990). Genetic flaxibility—a model for determining the range of suitable environment for a seed source. *Silvae Genet.* 39, 3–4.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rehfeldt, G. E., Ying, C. C., Spittlehouse, D. L., and Hamilton, D. A. Jr. (1999). Genetic responses to climate in Pinus contorta: niche breadth, climate change, and reforestation. *Ecol. Monogr.* 69, 375–407. doi: 10.1890/0012-9615(1999)069[0375:GRTCIP]2.0.CO;2

Roberds, J. H., and Namkoong, G. (1989). Population selection to maximize value in an environmental gradient. *Theor. Appl. Genet.* 77, 128–134. doi: 10.1007/BF00292327

Sas Institute Inc. (2012). *SAS OnlineDoc 9.3*. Cary, NC: SAS Institute Inc.

Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity. *Annu. Rev. Ecol. Syst.* 24, 35–68. doi: 10.1146/annurev.es.24.110193.000343

Van Tienderen, P. H., and Koelewijn, H. P. (1994). Selection on reaction norms, genetic correlations and constraints. *Genet. Res.* 64, 115–125. doi: 10.1017/S0016672300032729

Via, S., Gomulkiewicz, R., De Jong, G., Scheiner, S. M., Schlichting, C. D., and Van Tienderen, P. H. (1995). Adaptive phenotypic plasticity: consensus and controversy. *Trends Ecol. Evol. (Amst.)* 10, 212–217. doi: 10.1016/S0169-5347(00)89061-8

West, G. B., Brown, J. H., and Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature* 413, 628–631. doi: 10.1038/35098076

Yang, R.-C. (2007). Mixed-model analysis of crossover genotype-environment interactions. *Crop Sci.* 47, 1051–1062. doi: 10.2135/cropsci2006.09.0611

Yang, R.-C., Crossa, J., Cornelius, P. L., and Burgueño, J. (2009). Biplot analysis of genotype × environment interaction: proceed with caution. *Crop Sci.* 49, 1564–1576. doi: 10.2135/cropsci2008.11.0665

Yang, R.-C., and Ham, B. (2012). Stability of genome-wide QTL effects on malt α-amylase activity in a barley doubled-haploid population. *Euphytica* 188, 131–139. doi: 10.1007/s10681-012-0680-6

Yang, R.-C., Stanton, D., Blade, S. F., Helm, J., Spaner, D., Wright, S., et al. (2006). Isoyield analysis of barley cultivar trials in the Canadian Prairies. *J. Agron. Crop Sci.* 192, 284–294. doi: 10.1111/j.1439-037X.2006.00209.x

Yates, F., and Cochran, W. (1938). The analysis of groups of experiments. *J. Agric. Sci.* 28, 410, 269–288.

Zuo, W. Y., Moses, M. E., West, G. B., Hou, C., and Brown, J. H. (2012). A general model for effects of temperature on ectotherm ontogenetic growth and development. *Proc. R. Soc. B Biol. Sci.* 279, 1840–1846. doi: 10.1098/rspb.2011.2000

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 May 2014; accepted: 30 June 2014; published online: 22 July 2014.
Citation: Yang R-C (2014) Analysis of linear and non-linear genotype × environment interaction. Front. Genet. ***5****:227. doi: 10.3389/fgene.2014.00227*
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 21 July 2014
doi: 10.3389/fgene.2014.00225

A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects

Urko M. Marigorta* and Greg Gibson

Center for Integrative Genomics, School of Biology, Georgia Institute of Technology, Atlanta, GA, USA

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Rong-Cai Yang, University of Alberta, Canada
Chirag Patel, Harvard Medical School, USA

****Correspondence:***
Urko M. Marigorta, Center for Integrative Genomics, School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA
e-mail: urko.martinez@biology.gatech.edu

The switch to a modern lifestyle in recent decades has coincided with a rapid increase in prevalence of obesity and other diseases. These shifts in prevalence could be explained by the release of genetic susceptibility for disease in the form of gene-by-environment (GxE) interactions. Yet, the detection of interaction effects requires large sample sizes, little replication has been reported, and a few studies have demonstrated environmental effects only after summing the risk of GWAS alleles into genetic risk scores (GRSxE). We performed extensive simulations of a quantitative trait controlled by 2500 causal variants to inspect the feasibility to detect gene-by-environment interactions in the context of GWAS. The simulated individuals were assigned either to an ancestral or a modern setting that alters the phenotype by increasing the effect size by 1.05–2-fold at a varying fraction of perturbed SNPs (from 1 to 20%). We report two main results. First, for a wide range of realistic scenarios, highly significant GRSxE is detected despite the absence of individual genotype GxE evidence at the contributing loci. Second, an increase in phenotypic variance after environmental perturbation reduces the power to discover susceptibility variants by GWAS in mixed cohorts with individuals from both ancestral and modern environments. We conclude that a pervasive presence of gene-by-environment effects can remain hidden even though it contributes to the genetic architecture of complex traits.

Keywords: gene-by-environment, environmental perturbation, modern lifestyle, complex disease, genetic risk score, decanalization, GWAS, obesity

INTRODUCTION

Diseases such as diabetes, cardiovascular disease, and obesity have become highly prevalent in the developed world in a period of just a few generations. For example, more than one third of U.S Citizens are obese (Ogden et al., 2007). The incidence of these "modern" diseases is now also rising in developing countries (Abegunde et al., 2007). Recent changes in lifestyle are thought to be the main drivers of the emergence of these diseases, because genetic changes at the population level only occur after many generations.

Paradoxically, the rapid increase in prevalence of these diseases coincides with large heritability values. There is increasing evidence that the heritability of several traits has increased in the last 50 years. Obesity serves to illustrate this point. An analysis of Swedish military conscripts born from 1951 to 1983 showed an increase in the heritability along with a marked increase in the genetic variance for obesity (Rokholm et al., 2011b). A further study of Danish twins showed that one percentage point increase in the prevalence of obesity accompanies a ∼3.3% increase in the genetic variance for the trait (Rokholm et al., 2011a). Thus, the increased influence of the current "obesogenic" environment exerts its effects through a large alteration in the overall contribution of genetic factors to the susceptibility for obesity. The two most likely explanations for this phenomenon consist of (i) uncovering of new cryptic susceptibility variants that did not previously participate in the genetic architecture of the trait (Gibson and Dworkin, 2004), or (ii) an increase in the effect size of variants already associated with obesity before the emergence of the current "obesogenic" environment (Hermisson and Wagner, 2004).

In the last 5 years, thanks to the detection of genetic variants robustly associated by GWAS, the presence of gene-by-environment interactions (GxE) has been confirmed for several traits. However, the discovered GxE effects explain just a minor fraction of variance, suggesting that most interaction effects remain hidden. The poor availability of reliable environmental data constitutes one the major hurdles to detect GxE interactions. Genetic variation of common nature can be interrogated systematically with commercial genotyping arrays, but the availability of counterpart environmental information is often patchy and inconsistent, impeding a systematic interrogation of GxE effects (Patel et al., 2010, 2013). Moreover, the lack of high-throughput environmental data makes it difficult to replicate consistently GxE findings across datasets (Patel and Ioannidis, 2014). A second obstacle lies in the large sample size that is required to discover interaction effects univocally. For example, an early report observed that physical activity and diet modulate the effects of FTO variants on obesity (Demerath et al., 2011), but the evidence remained unclear in subsequent studies (Hubacek et al., 2011; Van Vliet-Ostaptchouk et al., 2012) until a large meta-analysis of 45 studies of ∼240,000 samples confirmed this interaction. Specifically, this meta-analysis established that

the risk effect of FTO alleles was ∼100 and 40% larger in physically inactive relative to active individuals from North America and Europe, respectively [Odds Ratio: 1.43 vs. 1.22 and 1.27 vs. 1.21, respectively (Kilpelainen et al., 2011)].

Additionally, synergistic interactions between causal alleles and environmental factors are being detected through genetic risk scores (Franks et al., 2013). The calculation of GRS involves generation of a weighted sum of the risk due to several variants into a single figure, thus overcoming the limitation of statistical power for individual SNPs. For example, the interaction between risk alleles and sugar-sweetened beverage intake has been confirmed by means of a predisposition score for obesity based on 32 GWAS-discovered obesity variants. Specifically, the risk in BMI per 10 risk alleles increased by 77% in individuals with ≥1 serving per day compared to sugary beverage intake <1 serving per month (Qi et al., 2012). Similar examples of GRSxE detection have been described for fried food consumption and adiposity (Qi et al., 2014), cigarette use polygenic risk and neighborhood social cohesion (Meyers et al., 2013) and Western dietary patterns and type 2 diabetes (Qi et al., 2009; Nettleton et al., 2013).

In order to quantify how prevalent this GRS-by-environment (GRSxE) contribution may be, we have performed a simulation study of a quantitative trait under "ancestral" and "modern" environments. Our main aim was to define the range of realistic conditions in which GRSxE interaction effects can be detected in the absence of evidence for individual GxE for the contributing alleles. The environmental perturbation and genetic architecture of the trait are based on recent inferences from human GWAS data. We demonstrate that a wide range of perturbation effects is consistent with current observations from GxE studies, although our investigations also show that these effects may heavily reduce the power to detect causal alleles by GWAS.

MATERIALS AND METHODS

GENETIC ARCHITECTURE OF THE SIMULATED TRAIT

We performed simulations of a polygenic quantitative trait to study the feasibility to detect gene-by-environment effects in the context of GWAS studies. We considered a trait partially controlled by genetic variants in the context of a total phenotypic variance of 1 ($V_P = 1$). In all simulations, we approximate the genetic architecture based on two recent inferences regarding the genetic basis of complex traits in humans. First, the trait is controlled by 2500 causal SNPs of common nature (minor allele frequency >5%). This number of genes resembles the number of susceptibility variants inferred for several complex traits [e.g., from ∼1700 to 2900 for myocardial infarction and type 2 diabetes, respectively (Stahl et al., 2012)]. Second, we assign the percentage of variance explained by each causal SNP (genetic variance of the trait, *gv*) based on the inferences from a large meta-analysis on normal height variation (Lango Allen et al., 2010). This study discovered 180 loci associated with height, each explaining from 0.012 to 0.28% of the variance in the trait. The contribution of 701 variants of similar effect size (accounting for 16% of the V_P) was inferred. We thus assigned the inferred distribution to 701 randomly selected variants from the 2500 simulated SNPs (gathered from Supplementary Table 4 in Lango Allen et al., 2010). Each of the remaining 1799 alleles was assigned to explain 0.012% of the variance. Hence, the 2500 simulated common SNPs individually explain from 0.012 to 0.28% of the variance, and the total genetic component of the trait accounts for 36% of the V_P (heritability = 36%). Importantly, note that we assign the allelic effects as a percentage of variance that each SNP explains, with the corollary that the actual effect size per allele will depend on the frequency of the causal allele (see next paragraph).

The number of SNPs and *g* of the trait are fixed. Then, for each simulation we re-assign the effect allele frequencies (EAF) and effect sizes (β) at each of the 2500 causal SNPs. To mimic the ascertainment bias of GWAS arrays, EAF values were drawn from a uniform distribution with boundaries 0.05 and 0.95 [$U_{(0.05,0.95)}$]. Genotypes were simulated assuming Hardy-Weinberg equilibrium. For example, for a SNP with EAF = 0.4 in a simulation of 10,000 samples, we would assign a value of 0, 1, and 2 phenotype-increasing alleles to ∼1600, 4800, and 3600 individuals, respectively. At this point of each simulation, we know the number of alleles that every individual carries at each site, as well as the total genetic variance each SNP explains. We can then easily calculate the effect size (β) of each SNP. The effect of the i^{th} SNP on the trait is given by its contribution to the genetic variance of the trait (Park et al., 2010):

$$gv_i = 2 * \beta^2 * EAF_i * (1 - EAF_i)$$

For example, a variant that explains 0.28% of the V_P with an effect allele frequency of 0.4 would increase the simulated phenotype by 0, 0.076, and 0.153 in individuals with 0, 1, and 2 causal alleles at that position, respectively. We consider an additive polygenic architecture. Thus, for each simulated individual the effects are added additively per allele copy, and summed independently across all 2500 causal loci. After assigning the effects to all SNPs, the additive genetic variance component (V_A) equals ∼0.36. To achieve the desired phenotypic variance ($V_P = 1$), we assigned a random environmental component (V_E) to every individual, drawn from a normal distribution with mean 0 and variance 0.64 ($V_E = 1 - V_G$). In summary, we simulated a quantitative trait with heritability 36% that results from the additive gene action over 2500 independent causal SNPs of common frequency.

MODELING A SHIFT IN ENVIRONMENT THAT PERTURBS THE GENETIC EFFECT SIZES

The genetic architecture explained above assumes that all individuals experience the same environment. This study investigates the consequences of a change in the environment that also modifies genetic contributions to disease or traits. Consequently, for convenience we call the baseline situation the "ancestral" environment, and postulate a new "modern" environment in which genetic effects are perturbed at some fraction of the 2500 causal SNPs. We also suppose that in contemporary society, some individuals have a lifestyle more close to the "ancestral" one (simplistically, low caloric intake, high activity) while others have a more "modern" lifestyle (they consume sugary beverages and engage in other obesogenic behaviors). In reality there will be a gradation, but the dichotomous model serves for purposes of illustration of the potential consequences for disease for contemporary societies of the transition to a western lifestyle, that may have induced GxE

effects (Gibson, 2009). Specifically, we considered the situation in which some or all individuals in the population live in a new environment that provokes a scaling effect (perturbation) in the genetic effect size at a fraction of the 2500 causal SNPs. Thus, simulated individuals can be classified into two binary "unperturbed" and "perturbed" categories, according to the environment they live in. The ancestral and modern environments aim to model a situation in which the genetic susceptibility to disease may have been altered in modern societies as a consequence of the transition to a western lifestyle (Gibson, 2009), that may have induced GxE through scaling effects. Specifically, the "modern" environment alters the genetic architecture of the trait by causing a multiplication of the effect size (β) by a constant factor (e.g., with a 1.5-fold change, a SNP with $\beta_{ANCESTRAL} = 0.06$ transforms to $\beta_{MODERN} = 0.09$). The strength of the GxE interaction is proportional to, first, the factor of perturbation and, second, the proportion of SNPs that become perturbed in the "modern" environment. For example, physical activity was shown to attenuate the association between rs9939609 in *FTO* and body mass index (BMI) by ∼30 to 95% (Andreasen et al., 2008; Kilpelainen et al., 2011). Another recent study on the interaction of sugar-sweetened beverages and BMI described an increase of 77% in the genetic risk per 10 causal alleles for individuals who drink >1 beverage serving per day, which would translate into an ∼8% increment in the effect size per variant under the "modern" environment (Qi et al., 2012). In our simulations, we explored the parameter space that ranges from 5 to 100% increase in the genetic effect size (1.05–2-fold change, respectively). Regarding the proportion of SNPs perturbed, we explored the outcomes after perturbing from a minimum of 1% to a maximum of 20% of the causal variants (25 and 500 of the 2500 simulated SNPs, respectively).

SELECTION OF SNPs THAT BECOME PERTURBED IN THE "MODERN" ENVIRONMENT

All causal SNPs do not account for the same proportion of genetic variance in the simulated trait. Therefore, the degree of GxE we induce also depends on the actual effect size of the perturbed SNPs. We explored two different models of SNPs that become perturbed. In model 1, the SNPs were chosen at random, whereas in model 2 they were chosen from those explaining most of the variance (e.g., the 250 SNPs with highest explained variance in simulations if 10% of the variants were perturbed). Importantly, the random environmental component (V_E) was drawn equally in both "ancestral" and "modern" environments. In other words, the "modern" environment induces an increase in the V_P after perturbation that is entirely dependent on the genetic component of the trait, thus increasing the V_G and heritability. Models entailing an increase in V_E could be similarly explored, but we do not do so here. Moreover, we note that although we only simulate scaling effects (at the SNP level), since only a small portion of variant effects is perturbed, there are also rank effects at the phenotype level.

THREE SCENARIOS OF SNP DISCOVERY IN A GWAS SETTING

For both perturbation models 1 and 2 explained above, we set up three different scenarios to perform a "SNP discovery" process to ascertain the variants that were subsequently tested for the presence of GxE effects (see a workflow summary in **Figure 1**). In the first scenario, "scenario A," we act as if all perturbed SNPs were known, and forward them directly to GxE analysis (see next section). "Scenario A" avoids the GWAS discovery step and thus constitutes an ideal situation to establish an upper bound for the range of perturbation effects that can be detected under models 1 and 2.

However, in reality we do not know in advance which SNPs may have undergone environmental perturbation in effect size. Usual practice consists on testing GxE effects for variants that have been previously associated by GWAS. To mimic the situation, we developed two further scenarios in which we added a preliminary GWAS step to discover SNPs. In "scenario B," we performed a GWAS in which 100% of the samples were selected from the "modern" perturbed environment. In "scenario C" we performed GWAS upon a sample that is drawn equally from each of the two environments (50% of the individuals come from the "ancestral" and "modern" settings, respectively). In other words, "scenario C" corresponds to a situation in which half of the society lives in an "ancestral" environment (e.g., extensive physical activity in daily life and low fat diet), whilst the other half follows a "modern" lifestyle that increases the effect size of perturbed alleles. Importantly, we do not "know" which environment each individual lives in, in the sense that this information is not included in the discovery GWAS. For both scenarios, we performed a two-stage genome-wide screen in which the quantitative phenotype is regressed against the allele dosage at each SNP. In the discovery screen, we assay the 2500 simulated SNPs in a sample of 50,000 individuals. SNPs that achieve $P < 10^{-5}$ in the discovery GWAS are then assayed in a meta-analysis with 100,000 individuals after joining the 50,000 samples from the discovery GWAS with a new simulated replication sample of 50,000 individuals. Finally, SNPs associated with the quantitative trait at $P < 5 \times 10^{-8}$ in the meta-analysis are then forwarded to a novel sample of 40,000 individuals for the GxE analysis described in the next section.

TESTING FOR GENE-BY-ENVIRONMENT EFFECTS AFTER PERTURBATION

A central focus of our study lies in the evaluation of the power to detect the GxE effects in our simulated trait. We aimed to evaluate the performance of two different approaches, namely (i) power of detection through the examination of individual SNPs and (ii) by means of unweighted genetic risk scores (GRS) that sum up the number of causal alleles for each individual (without weighting each allele by its effect size). To do so, for each scenario we simulated two cohorts of 20,000 individuals each that are sampled from the "ancestral" and "modern" environments, respectively. For each simulated individual, we know its phenotype, the number of causal alleles at each SNP (coded as "0," "1," and "2"), the total number of causal alleles over all selected loci (GRS) and the environment it belongs to (coded as "0" and "1" for "ancestral" and "modern" environments, respectively). In each simulation of 40,000 individuals, we tested the interaction between genetic component and environment by means of a multiple linear regression: $Y_j = \beta_0 + \beta_G * \chi_{(G)} + \beta_E * \chi_{(E)} + \beta_{(G*E)} * \chi_{(GE)}$

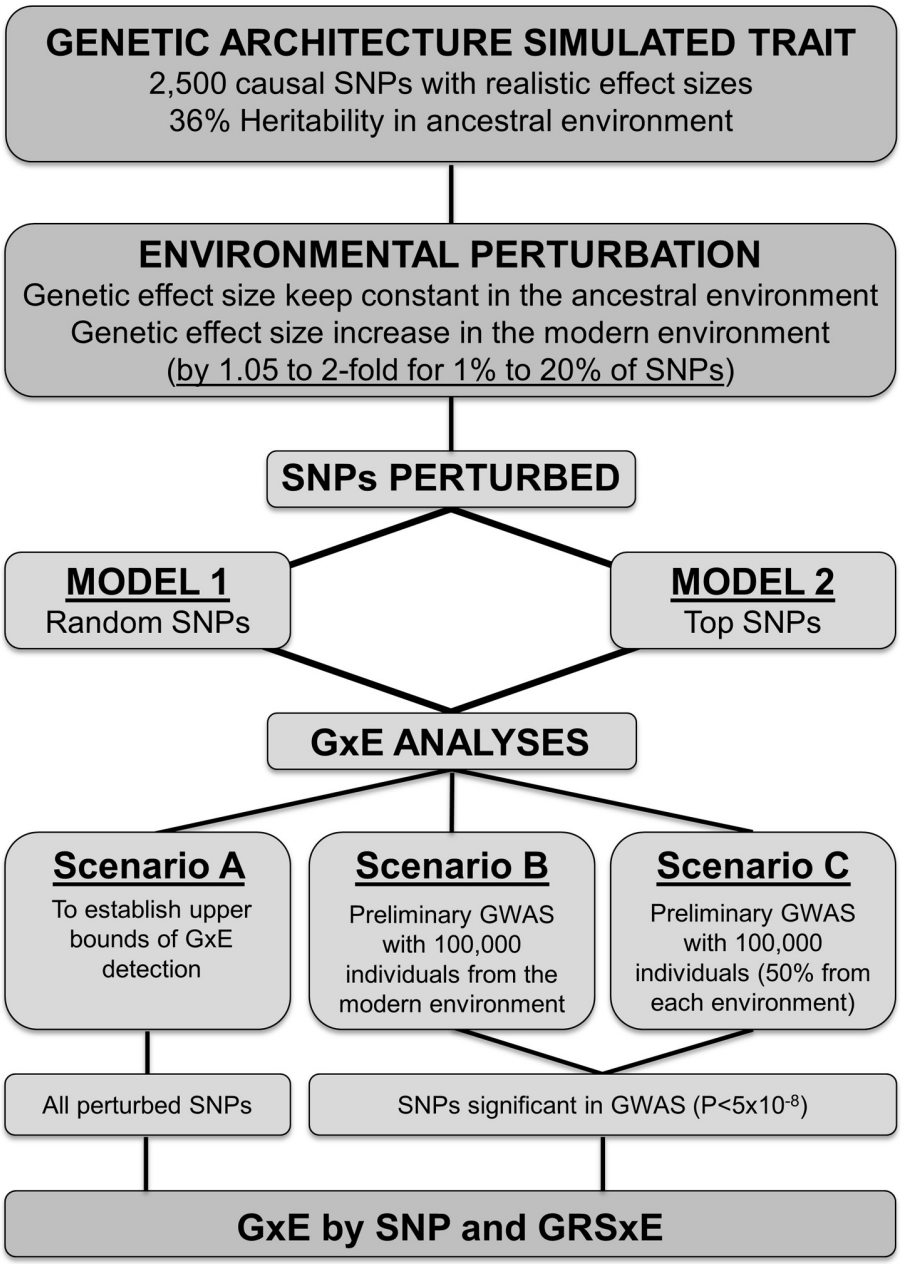


FIGURE 1 | Summary of the steps followed in the study.

to estimate the regression coefficient $\beta_{(G*E)}$, with Y_j, $\chi_{(G)i}$, and $\chi_{(E)i}$ recording the phenotype, allele dosage (or GRS) and environment of the individual j, for individuals 1, . . . , 40, 000.

In summary, we explored two different ways to select SNPs that undergo perturbation and three different procedures to choose the actual variants upon which we test for gene-by-environment interactions. For each of the six resulting combinations (models 1 or 2, and scenarios A, B, or C), we explored 400 combinations of parameters. Specifically, the percentage of SNPs that experienced perturbation ranged from 1 to 20% (20 steps of 1%), and the factor of perturbation ranged from a 1.05–2-fold change in effect size (20 steps of 0.05-fold increments). We performed five different replications for each of the 400 combinations, and thus 2000 simulations for each of the six combinations. Results are summarized as heat maps that interpolate relevant parameters across a continuous range of values (**Figures 2**, **4–7**, and Supplementary Table 1).

STATISTICAL ANALYSIS

All the analyses were performed using the R software v.3.0 (R Core Team, 2013). Associations between the simulated phenotype and allele dosage, as well as the GxE interactions, were tested with the

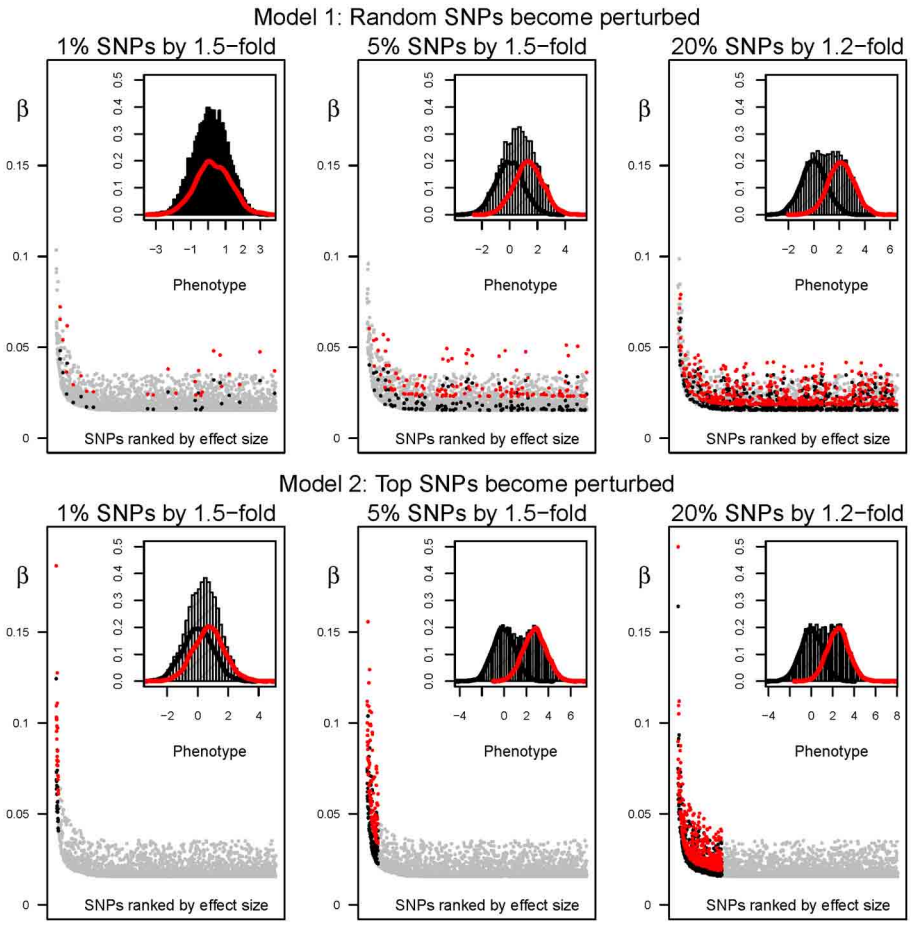


FIGURE 2 | Distribution of effect sizes and phenotypes before and after perturbation. From left to right, outcomes after perturbing the effect size of 1, 5, and 20% of the SNPs by 1.5, 1.5, and 1.2-fold, respectively. Top and bottom panels have been simulated under models 1 and 2, respectively. Major plot: the effect size (β) of the 2500 causal SNPs is shown in y-axis. SNPs are ranked in x-axis according to the percentage of variance explained by each variant in the "ancestral" environment (from left to right). Gray dots correspond to the effect size in the "ancestral" environment. The scaling effect among environments is shown in black and red. Specifically, the black and red dots in main panel indicate, respectively, the effect size before and after (in "ancestral" and "modern" environments). Minor plot: distribution of phenotypes in a cohort of 10,000 simulated individuals. Black and red curves show the histogram of phenotypes for the same individuals, before and after perturbation. The curve in the background represents the histogram of phenotypes if the two simulated samples are joined into a cohort of 20,000 individuals.

lm function. Heatmap plots were generated using the *fields* and *akima* R packages.

RESULTS

We simulated an environmental perturbation in genetic effect sizes to explore the feasibility of detecting gene-by-environment interactions. In the "ancestral" environment, the 2500 causal variants explained from 0.012 to 0.28% of the phenotypic variance. In the "modern" setting a percentage of variants ranging from 1 to 20% underwent perturbation, and their effect sizes increased by a constant factor that ranged from 1.05 to 2-fold. We applied two different models to select the causal SNPs that become perturbed in the second "modern" environment, and built three scenarios to select the SNPs upon which we investigated the feasibility of detecting gene-by-environment interactions following the workflow in **Figure 1**. A detailed summary of the results for each simulation is available in Supplementary Table 1.

EFFECTS OF THE "MODERN" ENVIRONMENT IN THE DISTRIBUTION OF EFFECT SIZE AND PHENOTYPES

The actual effect size of each causal allele depends on the frequency and variance explained by the causal variant. For example, we set the strongest contribution in the "ancestral" environment at ∼0.3% of the variance explained. If that allele has a frequency of 0.5, it would present an effect size of 0.075 (β_{ANC}), increasing the phenotype by 0, ∼0.075 and 0.15 in individuals with zero, one and two causal alleles, respectively. If it becomes perturbed in the "modern" environment by the strongest perturbation possible (2-fold change; $\beta_{MOD} = 2 * \beta_{ANC}$), the effect size would increase from ∼0.075 to 0.15. Thus, the variant would increase by 4-fold the percentage of phenotypic variance it accounts for, hiking from ∼0.3 to 1.2% (see Materials and Methods).

The differences in the distribution of phenotypes under each environment not only depend on the strength but on the proportion of variants that become perturbed in the "modern"

setting. The same perturbation inducing a 2-fold increment in the effect size, but acting upon 20% of the SNPs, would result in a distribution of phenotypes that do not overlap extensively. We illustrate the resulting distribution of phenotypes under the "ancestral" and "modern" environments for a range of perturbation effects in **Figure 2** (black and red lines, respectively). For instance, the average phenotype under "modern" conditions after perturbing 20% of the causal SNPs by 1.2-fold is two standard deviations above the average phenotype under the "ancestral" environment. Overall, perturbation leads to a flattened distribution of phenotypes when individuals from both environments are combined, and the increase of phenotypic variance is proportional to the percentage of people that live in the "modern" environment. The differences are strengthened under model 2, because the SNPs that already present the largest effect sizes in the "ancestral" environment are chosen for perturbation in the "modern" setting. Indeed, the most extreme simulated perturbation, such as multiplying the effect of 20% of the variants by two, results in bimodal distributions that can be easily distinguished and are probably biologically unrealistic. However, the differences are much subtler for most of the parameter space, and in next sections we refer to the parameter space that results in a change in the distribution of phenotypes that resembles that of typical traits such as contemporary BMI (see **Figure 3** for a real example based on the change in BMI shown by North American males).

The perturbation in genetic effect sizes prompted by the "modern" environment leads to an increase in the heritability of the quantitative trait (**Figure 4**). The phenotype presents a basal heritability of 36% in the "ancestral" environment, but it easily boosts in the "modern" setting. For instance, a 1.2-fold increase in the effect size of 20% of the causal SNPs leads to a heritability of ∼80%, and a similar effect is achieved with a 1.3 and 1.6-fold change acting upon ∼10 and 5% of the causal variants, respectively. This happens because the "modern" environment induces a hike in V_P that is entirely due to a higher V_G (we keep V_E constant, see Material and Methods). Again, the effect is more pronounced under Model 2 (**Figure 4B**). For instance, 2-fold increments in the effect size inevitably lead to unrealistic heritability values above 90% in the "modern" environment.

We illustrate the effects of the "modern" environment on (i) the genetic effect sizes of perturbed SNPs (major graphs in **Figure 2**), (ii) the differences in the distribution of phenotypes between the "ancestral" and "modern" lifestyles (small graphs in **Figure 2**), and (iii) the heritability of the simulated trait (**Figure 4**). We next describe the ability to detect gene-by-environment interaction effects induced by the "modern" setting. We compare the ability to detect GxE interactions at the SNP level with that of GRSxE analyses. Overall, we consider three different scenarios to ascertain candidate SNPs, and examine for GxE effects in cohorts of 40,000 individuals in which half of the samples come from the "ancestral" and "modern" environments, respectively.

DETECTION OF GxE EFFECTS WHEN ALL PERTURBED VARIANTS ARE KNOWN (SCENARIO A)

Even if the analyses include all variants that are perturbed (that is, known from the model, without a GWAS discovery step), GxE effects tend to remain undetected at the SNP level (see **Figure 5**). Specifically, under Model 1 only 32 out of 2000 simulations (1.6%) achieved genome-wide significance ($P < 5 \times 10^{-8}$) for any SNP in the GxE analyses, and all of these required a >1.5-fold change in the effect size (**Figure 5B**). Indeed, at most a single

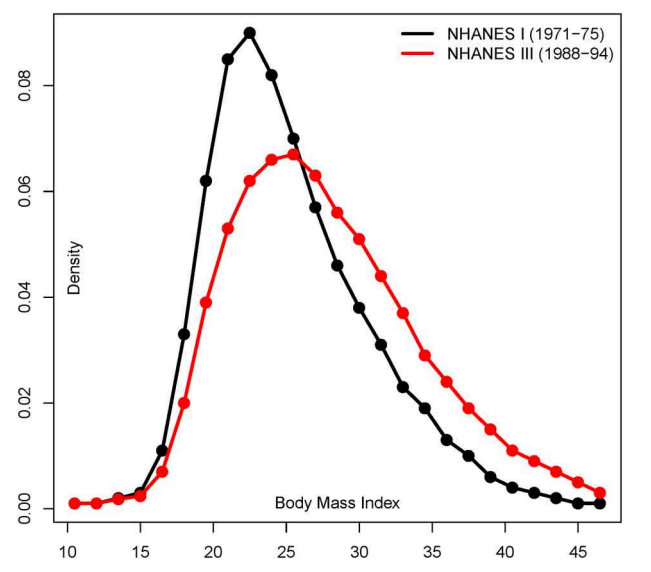


FIGURE 3 | Shift in BMI in U.S males from 1971–1975 to 1988–1994. Distribution of BMI in North American males (20–55 age) studied in the NHANES I and III health and nutritional surveys (adapted from Figure 1 in Cutler et al., 2003).

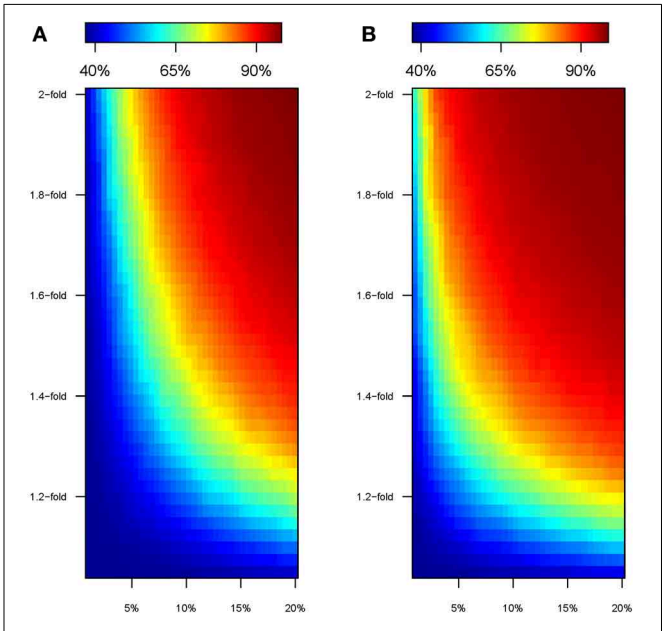


FIGURE 4 | Heritability of the simulated trait in the "modern" environment. Color map showing the heritability in cohorts perturbed under model 1 **(A)** and model 2 **(B)**, according to the percentage of SNPs perturbed (x-axis) and the factor of perturbation in effect size (y-axis).

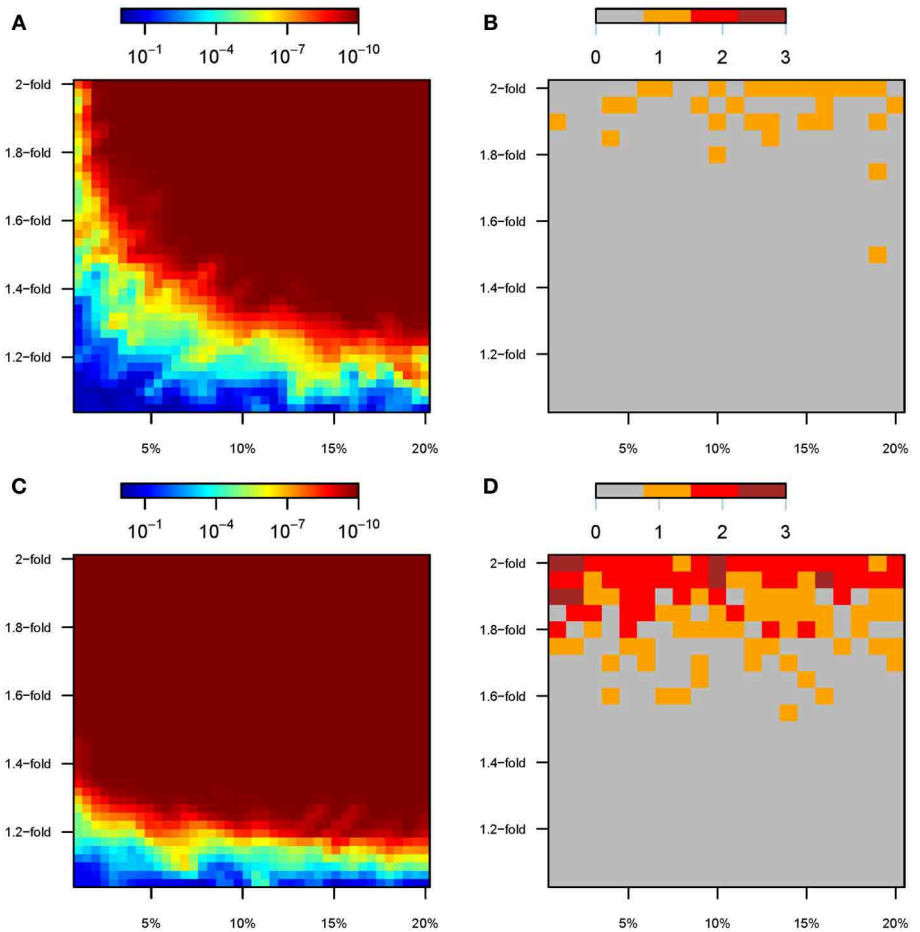


FIGURE 5 | GxE analyses under scenario A. For scenario A, color maps showing the results of the gene-by-environment interaction analyses according to the percentage of SNPs perturbed (x-axis) and the factor of perturbation in effect size (y-axis). **(A)** *P*-value of the GRSxE interaction under model 1. **(B)** Number of SNPs at genome-wide significance levels ($P < 5 \times 10^{-8}$) for GxE under model 1. **(C)** *P*-value of the GRSxE interaction under model 2. **(D)** Number of SNPs at genome-wide significance levels ($P < 5 \times 10^{-8}$) for GxE under model 2. Panels **(B,D)** record the largest number observed out of five permutations.

variant was detected in each simulation, even if we tested for GxE individually for all perturbed SNPs (e.g., 500 tests for GxE when 20% of the variants were perturbed). Furthermore, only 14% of the 100 simulations with a 2-fold change in the effect size harbored a variant that passed the threshold for genome-wide significance (**Figure 5B**). Conversely, there was a wide range of perturbation parameters for which genetic risk scores, the sum of the total number of causal alleles each individual carries, constituted a powerful tool to detect interaction effects induced by the "modern" environment (**Figure 5A**). For instance, GRSxE interaction terms using GRS calculated over 250 perturbed SNPs (10% of causal variants) showed extremely low *p*-values ($P < 10^{-10}$) for all the ranges from 1.3 to 2-fold change in the genetic effect size. Indeed, tiny increments in the effect size, such as a 1.2-fold change, resulted in ∼100% of the simulations detecting GRSxE effects at the $P < 0.05$ significance level (notice that we performed a single GRSxE test per simulation, because the allelic count of all tested variants were collapsed into a single number). Only the parameter space correspondent to <1.1-fold changes for <5% of the causal variants consistently resulted in non-significant GRSxE interaction terms (**Figure 5A**).

The same patterns were observed under the environmental perturbations of Model 2, although an overall increased ability to detect interaction effects was noticed (**Figures 5C,D**). Specifically, 12.8% of the simulations (255 out of 2000) led to significant GxE effect at the SNP level, although 74.1% of those showed a single variant being genome-wide significant (189 out of 255). It was necessary to perturb genetic effects by 1.8–2-fold to achieve several variants being significant at the SNP level (**Figure 5D**). The interaction effects induced by the "modern" environment are almost universally detected through GRSxE analyses (**Figure 5C**).

DETECTION OF CAUSAL ALLELES BY GWAS AFTER MODERN ENVIRONMENTAL PERTURBATION

In "scenario A," the environmental perturbation in effect sizes can be easily detected with GRSxE analyses. These results establish

an upper bound for the ability to detect gene-by-environment effects induced by the "modern" lifestyle, because the analyses are restricted to the truly perturbed variants. Yet, for real traits it is uncertain which SNPs may present GxE effects. Usual practice consists of prioritizing variants unequivocally associated to the trait of interest, such as the alleles discovered by GWAS. To mimic this procedure, we perform a preliminary GWAS study to ascertain variants for GxE analyses.

GWAS meta-analyses of 100,000 individuals entirely drawn from the "ancestral" environment detected ∼90 genome-wide significant variants, accounting for ∼15% of the heritability (data not shown). GWAS on cohorts with 100% of the individuals being "perturbed" under model 1 led to an increased ability to detect variants associated to the trait (**Figure 6A**). The number of detected variants oscillated from 100 to 150 for the most realistic range of perturbation parameter space, and hiked to ∼300 when GWAS was performed upon 100,000 very heavily "disturbed" individuals (e.g., 2-fold change in the effect size for ∼20% of the causal variants). A progressively larger number of the associated variants that are detected correspond to perturbed variants (**Figure 6B**). The tendency to detect increasing proportions of perturbed variants becomes exacerbated under model 2. Specifically, and even if similar numbers of significant variants are detected by GWAS (**Figure 6C**), the increment in SNP detection corresponds to perturbed variants (**Figure 6D**).

Highly divergent patterns were observed when we perform a preliminary GWAS upon a mixed sample of individuals drawn equally from the "ancestral" and "modern" environment ("scenario C"). Under Model 1, the number of variants detected by GWAS still remained close to ∼90 only if the 50% of GWAS individuals coming from the "modern" environment had only been perturbed slightly (e.g., <1.2-fold for <5% of the causal SNPs, bottom-left corner in **Figure 6E**).The ability to detect causal alleles dropped when more extensive perturbations were simulated. For instance, ∼60 variants were detected at genome-wide significance levels when 7% of the variants had their effect size multiplied by 1.3-fold, and almost no variants are discovered if the same percentage of SNPs underwent a 1.8-fold change in

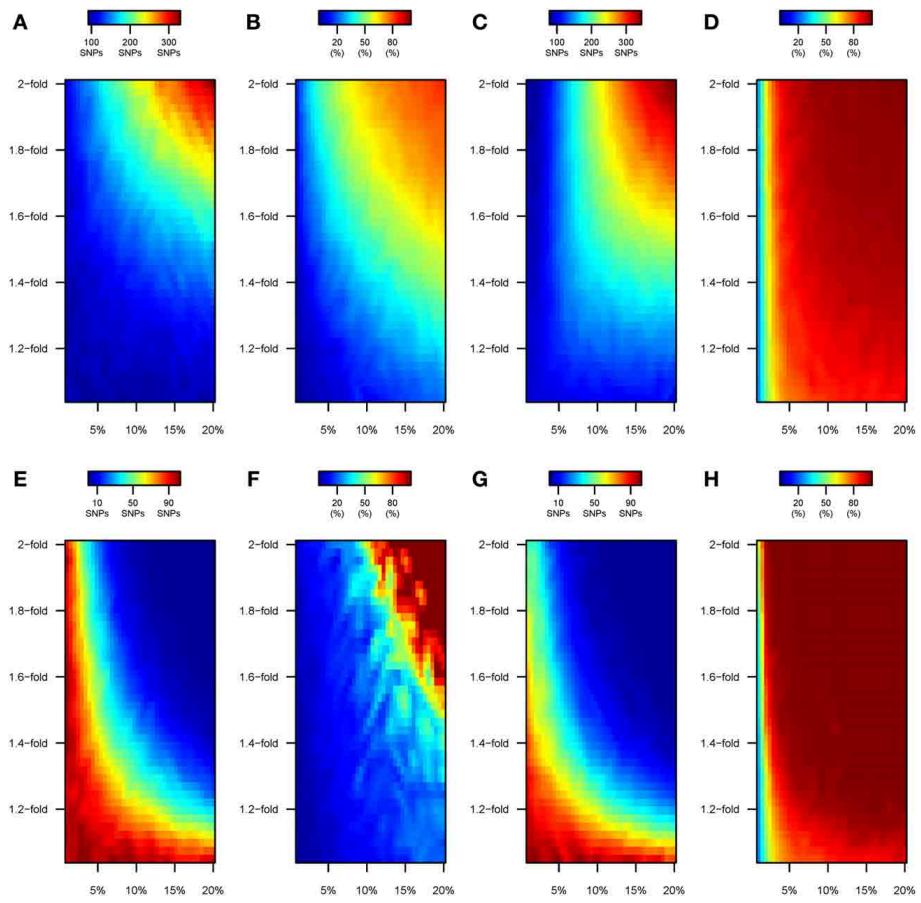


FIGURE 6 | Number of SNPs discovered by GWAS under scenarios B and C. Color maps showing the results of the GWAS upon cohorts of 100,000 individuals with (i) 100% of the samples drawn from the "modern" environment (scenario B; top panels, **A–D**) and (ii) 50% of the samples drawn from each "ancestral" and "modern" environments (scenario C; bottom panels, **E–H**). Specifically: **(A,E)** Under model 1, number of variants discovered by GWAS at genome-wide significance levels ($P < 5 \times 10^{-8}$). **(B,F)** Under model 1, percentage of the genome-wide significant variants that have undergone perturbation. **(C,G)** Under model 2, number of variants discovered by GWAS at genome-wide significance levels ($P < 5 \times 10^{-8}$). **(D,H)** Under model 2, percentage of the genome-wide significant variants that have undergone perturbation.

the effect size, or with a 1.3-fold increase for 20% the causal SNPs. Interestingly, the increasingly reduced number of variants discovered by GWAS under "scenario C" corresponded to perturbed SNPs (top-right corner in **Figure 6F**). Similar patterns were observed for "scenario C" under model 2 of perturbation (**Figures 6G,H**). As discussed below, we attribute these effects to the increase in phenotypic variance being greater than the individual genetic effects of each SNP.

DETECTION OF GENE-BY-ENVIRONMENT INTERACTIONS WITH SNPs DETECTED BY GWAS

The enhanced power to discover SNPs under "scenario B" resulted in patterns of GxE interaction detection that are similar to those observed for "scenario A," in which only perturbed variants were used (**Figures 7A–D**). SNP-by-SNP tests rarely resulted in significant GxE interaction coefficients (**Figure 7B**). By contrast, a wide range of the parameter space led to significant GRSxE evaluations, starting from ∼1.4-fold change for ∼5% of the variants to any stronger perturbation, **Figure 7A**). Similarly, under model 2 the tendency toward significant GRSxE detection was exacerbated (**Figure 7C**), and GRSxE interactions were significant for the whole range of simulated parameters. In these analyses, only GWAS performed upon strongly perturbed individuals (1.8–2-fold change in β) permitted detection of perturbed SNPs that were consistently significant at the individual level in the GxE analysis (**Figure 7D**).

A reversed pattern was observed under "scenario C." The proportion of perturbed SNPs among the detected variants was higher as perturbation strengthened, but it became negligible in absolute terms because almost no variants were detected by GWAS. Thus, the overall poor performance of mixed GWAS to detect perturbed SNPs rendered almost impossible the task of detecting GxE effects with GWAS SNPs, even at the GRSxE level (**Figures 7E,F**). The compromised detection power under "scenario C" does not however preclude the detection

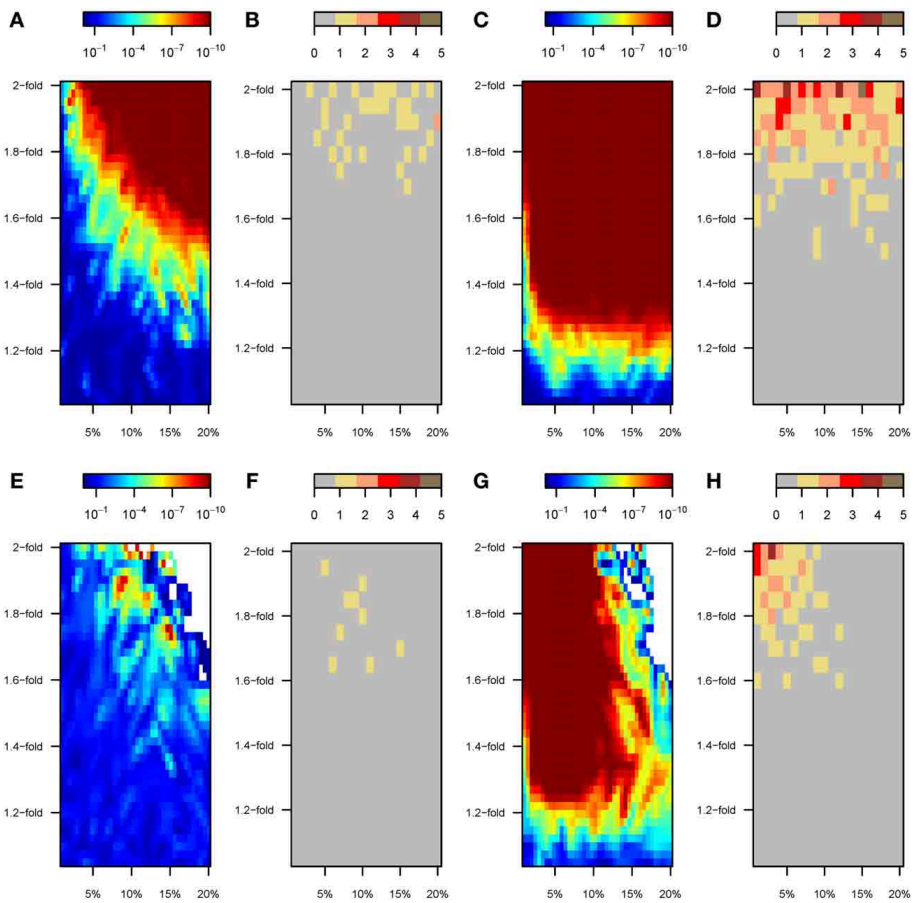


FIGURE 7 | GxE analyses with SNPs discovered in a preliminary GWAS (scenarios B and C). Color maps showing the results of the gene-by-environment interaction analyses according to the percentage of SNPs perturbed (x-axis) and the factor of perturbation in effect size (y-axis). Results for scenario B are shown in top panels **(A–D)**. Specifically: **(A)** *P*-value of the GRSxE interaction under model 1. **(B)** Number of SNPs at genome-wide significance levels ($P < 5 \times 10^{-8}$) for GxE under model 1. **(C)** *P*-value of the GRSxE interaction under model 2. **(D)** Number of SNPs at genome-wide significance levels ($P < 5 \times 10^{-8}$) for GxE under model 2. The corresponding results for scenario C are shown in bottom panels **(E–H)**. Panels **(B,D,F,H)** record the largest number observed out of five permutations. White areas in top right corners in panels **(E,G)** correspond to parameter space with no SNPs detected by GWAS and thus missing GRSxE analyses.

of gene-by-environment effects through GRSxE analyses under model 2 (**Figures 7G,H**).

DISCUSSION

In this study we performed a series of simulations to inquire under what conditions gene-by-environment effects can be detected. We applied an environmental perturbation upon cohorts of individuals that live in either an "ancestral" environment, or a "modern" setting that leads to an increment in the genetic effect sizes of a percentage of the causal alleles. For a wide range of the explored parameter space, gene-by-environment effects mostly remain unnoticed when interaction is examined at the SNP level. Conversely, GxE analyses are well powered to detect significant interactions when the genetic component of each individual is summarized through genetic risk scores (GRS) that sum up the total number of causal alleles in a single figure. Moreover, we find that the ability to detect perturbed SNPs in a GWAS preliminary to the GxE analysis depends on the mixture of samples coming from each environment. Genome-wide screens performed upon homogeneous cohorts of perturbed individuals show increased power to detect significant gene-by-environment interaction effects. In contrast, GWAS upon heterogeneous mixtures of "unperturbed" and "perturbed" individuals present a decreased ability to detect significant SNPs, thus inhibiting the task of detecting GxE effects.

FEASIBILITY OF THE ENVIRONMENTAL PERTURBATION UNDER THE "MODERN" ENVIRONMENT

The validity of the insights gained from this study depends on the plausibility of our model of environmental perturbation, and the extent to which we mimic the reality faced by current GWAS studies. Certainly, it is difficult to evaluate the consequences of the "modern" perturbation in the case of actual human phenotypes because the heritability and phenotype distributions correspondent to the "ancestral" lifestyle are unknown. However, there is increasing evidence that the switch to a western lifestyle may have been coupled with a change in the genetic effects of causal alleles (Gibson, 2009). Human complex traits result from the assemblage of multiple physiological dimensions, which may lead to a canalization of phenotypes whereby genetic effects are minimized following long-term stabilizing selection (McGrath et al., 2011). Under such a theoretical model, the "modern" human standard of living may have uncovered the activity of previously silent, or almost silent, cryptic genetic variability (Hermisson and Wagner, 2004). For example, this could have been the case for polymorphisms lying in genes that participate in pathways involved in neural regulation of appetite (Heber and Carpenter, 2011). These variants may have played a small role in the genetic etiology of weight throughout the history of our species, but may explain a larger proportion of the individual susceptibility to obesity in the modern environment of unrestricted access to processed food. A variety of other similar situations could be imagined, such as the interplay between addiction, tobacco use and lung cancer (Amos et al., 2008). In our simulations, we explore a range of parameter space in which the "modern" environment perturbs from 1 to 20% of the causal variants. Such a change can be easily framed in a pathway perspective. Specifically, one or several physiological pathways participating in the genetic architecture of complex traits may respond differently under the "modern" environment. In the context of a common disease, the environmental perturbation we explore would plausibly amount to an increase in the proportion of the population at risk (as in **Figure 3** for real BMI).

Our model postulates one of the simplest instances of GxE in which individuals are assigned to a binary environmental state that would roughly correspond to "ancestral" and "modern" lifestyles. A more realistic scenario of environmental perturbation should summarize the varying fraction of "modern lifestyle" followed by each person into an individual-specific measure, or "exposome" (Patel and Ioannidis, 2014). More complex simulations could be tuned to incorporate more realistic settings. For instance, the extent of exposure to modern lifestyle could be more finely determined (e.g., degree of sedentary behavior, diet patterns, stress at work. . .) to explore threshold-dependent models of GxE. Our simulations are necessarily a simplification of the almost infinite array of GxE interactions that could arise in the presence of multi-layered and continuous environments that can perturb the genetic effects of causal variants (Luan et al., 2001; Wong et al., 2003). However, the qualitative environmental states in our simulations resemble the practice of recent studies that have confirmed GxE effects after categorizing the environment into binary categories, as has been the case for example in studies of sugar-sweetened beverage consumption and overall diet patterns (Do et al., 2011; Qi et al., 2012).

In addition to the mechanism of perturbation and the binary nature of the simulated environment, the realism of our perturbation model also depends on the likelihood that the explored parameter space is realistic. We chose to approximate this by checking whether the range of simulated effects results in phenotypic distributions that approximate real observations. In the context of BMI, for instance, western urban women have been shown to present an average BMI value that is ∼4 standard deviations larger than the corresponding figure for Hadza hunter-gatherer women (see Table 1 in Pontzer et al., 2012). These differences are similar to the average horizontal shift between "ancestral" and "modern" environment that we observe in our simulations (e.g., depending on the percentage of perturbed SNPs, changes in effect sizes by <1.4-fold lead to ∼1 to 4 standard deviations of difference in the average phenotype). Furthermore, we also examined the shape of the phenotype distributions. Indeed, we observe significant GRSxE analyses for simulations with parameter combinations that result into more flattened but unimodal distributions of phenotypes, such as those observed in U.S men (**Figure 3**). Nonetheless, the actual phenotypic variance of a combined population depends on the mixture proportions and even extreme situations in which half of the individuals are raised in each environment do not lead to a bimodal phenotypic distribution in a combined simulation population. The heritability of the trait is also kept within a reasonable range. It can severely hike to 90% in the context of the most severe perturbations, but the actual heritability would lie from 36 to 80% according to the exact proportion of "unperturbed" and "perturbed" individuals.

DETECTION OF GENE-BY-ENVIRONMENT EFFECTS WITH GENETIC RISK SCORES

We observe a substantial parameter space in which gene-by-environment effects can be easily detected with genetic risk scores while remaining hidden in individual SNP analyses, even after testing exclusively those variants that were detected in populations perturbed by the "modern environment." SNP-by-SNP analyses provide anecdotal evidence for significant GxE, and only when extreme perturbations are assayed (e.g., >400 SNPs perturbed by 2-fold in the effect size are necessary to detect a single genome-wide significant variant). Conversely, GRSxE analyses are always significant when β-s are multiplied by 1.3-fold or more, or for the whole range of perturbation parameters when the "modern" environment affects the SNPs that explain most of the variance in the trait (i.e., model 2). These results confirm that a widespread presence of GxE effects is not at odds with the lack of evidence when individual variants are assayed, despite of a substantial presence of interaction effects.

An important aspect of our simulations lies in the choice of variants that are perturbed by the "modern" environment. We observe that it is easier to detect GxE effects when the variants that are perturbed coincide with the alleles that explain most of the genetic basis of the trait, as in model 2. This makes sense considering that these perturbed variants not only present the largest effect sizes, but also have multiplied it in the "modern" environment. The same mechanism explains the increment in the number of variants detected by GWAS when the genome-wide screen is performed entirely upon perturbed individuals, as in "scenario B." For real traits with widespread GxE effects, it may be key to perform GWAS selecting for perturbed individuals. The selection of those individuals following a "modern" lifestyle would unravel specific pathways that respond badly in face of perturbation, thus enabling a more detailed understanding of the etiology of the diseases of affluence. Nonetheless, it may be inherently complex to design "perturbed-only" GWAS, owing to the difficulty in defining what exactly constitutes the perturbed environment. The sampling of individuals could also be confounded by the fraction of cases that are entirely due to purely environmental causes without any major role of gene-by-environment interactions linked to "modern" life.

MIXTURE OF ENVIRONMENTS COMPROMISES GWAS DISCOVERY POWER

The simulations in which the preliminary GWAS is performed upon cohorts with a mixed environmental exposure ("scenario C") show a remarkable trend regarding SNP discovery. The combination of "ancestral" and "modern" environments does not compromise the detection of causal variants when perturbation effects are tiny or restricted to a small fraction of the causal SNPs. However, larger perturbations decrease the ability to detect new variants, and statistical power eventually collapses for the strongest range of effects in our simulations. This result makes sense because gene-by-environment interactions add variance and heterogeneity in the estimates of SNP effects. We show the results for a causal variant that explains 0.3% of the variance in an "ancestral" population (**Figure 8**). This allele achieves $P < 10^{-12}$ when assayed in a GWAS with 20,000 individuals that follow the "ancestral" lifestyle. In contrast, the significance worsens ($P < 10^{-7}$) when this variant is assayed upon a mixture of 10,000 "ancestral" individuals and 10,000 individuals in which 10% of the SNPs have increased their effect size by 1.5-fold. Eventually, the variant remains completely unnoticed in a mixed GWAS when the effect size increases by 2-fold in the individuals following "modern" lifestyle ($P \sim 10^{-4}$). As a consequence, these variants are not found among the top candidate list in our simulated meta-analysis GWAS.

It is difficult to evaluate the extent to which pervasive gene-by-environment effects have compromised the power to discover associated variants by GWAS. The number of discovered variants correlates with sample size (Visscher et al., 2012), but some other differences among studies can be remarked upon. For instance, a large meta-analysis of ∼180,000 individuals reported 180 different loci associated to height, whereas a similarly powered study with >250,000 individuals only described 32 loci for BMI (Lango Allen et al., 2010; Speliotes et al., 2010). This may be explained simply by a difference of narrow sense heritability. On the other hand, the SNP-based heritability in these studies explains a notably greater proportion of the total heritability for height, implying a reduced missing heritability concern. We propose that this difference might be attributed to environmentally-induced heterogeneities in genetic effect size being more prevalent in the case of BMI, in turn explaining the lack of power to detect obesity-related loci. Arguably, this limitation can be avoided in real GWAS through the inclusion of covariates (e.g., variables that capture nutrition and physical activity levels per individual in a GWAS for obesity). However, the potential covariates to be included are often unknown or not available for all the cohorts, as in for example the largest meta-analyses for height and BMI (Lango Allen et al., 2010; Speliotes et al., 2010).

We explore a genetic architecture and a range of perturbation parameters that are based on empirical data, which strengthens the validity of our observations. However, the present study is not devoid of weaknesses. Among others, we have used the same sample size in all the simulated GWAS and GxE studies. This comes at a price, since the range of perturbations that result in significant GRSxE would certainly change if larger studies were performed. Second, we performed simulations of random mating populations with genotypic proportions following strict Hardy-Weinberg equilibrium (HWE). This procedure follows the usual practice consisting of screening polymorphisms for HWE. Nonetheless, confounding of population structure with environmental variability, further complicating the detection of GxE in real studies, remains a possibility. Third, we explored the presence of interactions through unweighted GRS that do not take into account the effect size of each variant. Since only a few variants present notably large effects (**Figure 2**), in reality weighted and unweighted risk scores are very highly correlated once more than a few dozen loci are incorporated, which minimizes the loss of power to detect GRSxE effects compared to weighted risk scores. Finally, it should be noted that we only simulate causal variants instead of tagging SNPs, which effectively over-estimates effect sizes relative to those discovered in true GRS.

In summary, the present study constitutes a preliminary evaluation of a realistic mechanism by which gene-by-environment

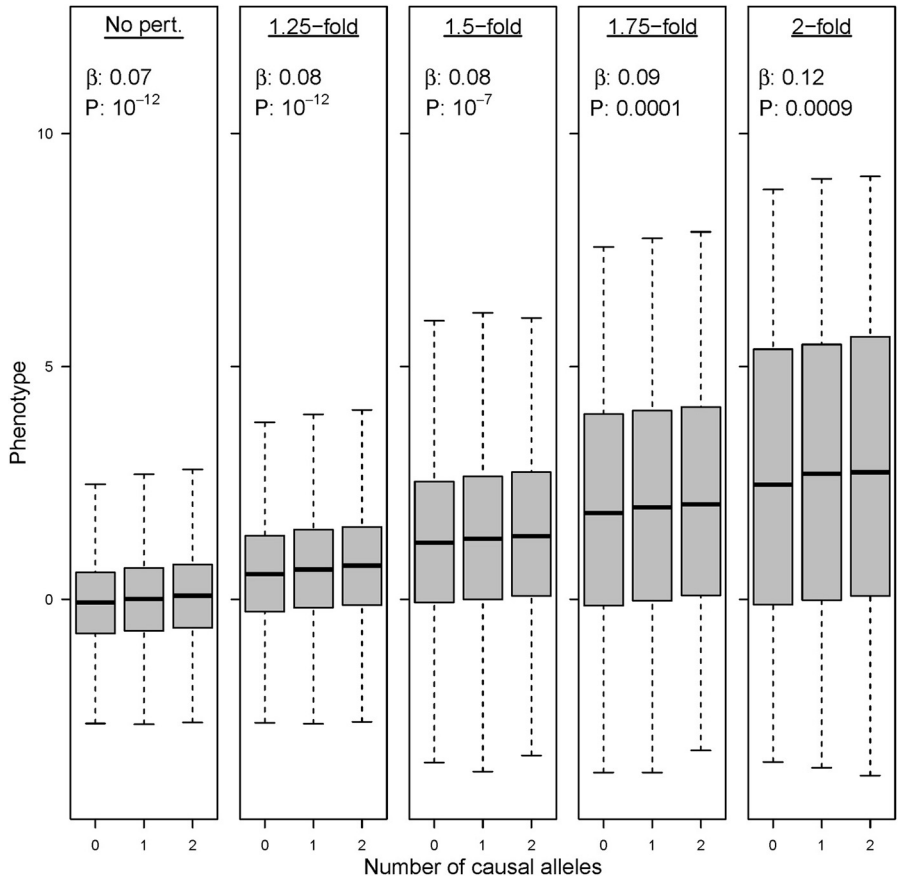


FIGURE 8 | Environmental perturbation in genetic effect sizes decreases the power of GWAS. Association results and *P*-value for the same variant under five different GWAS with 20,000 individuals. Left boxplot: a variant explaining 0.3% of the phenotypic variance achieves genome-wide significance in a GWAS with 100% of the samples being drawn from the "ancestral" environment. Successive boxplots: the same variant drops in statistical significance when tested in GWAS in which the allele has undergone a 1.25, 1.5, 1.75, and 2-fold perturbations in 50% of the individuals.

interactions may have altered the genetic etiology of human traits. A widespread presence of realistic GxE effects could only be detected by genetic risk scores calculated upon all variants discovered by GWAS. The extent to which these effects have shaped real human traits remains as an open question, and should be studied in future research.

AUTHOR CONTRIBUTIONS

GG conceived the original idea. UMM and GG designed the study. UMM performed the simulations. UMM and GG interpreted the data and wrote the paper.

ACKNOWLEDGMENTS

We acknowledge Kevin Lee and other colleagues from Gibson's lab and Isabel Mendizabal for their helpful comments during this work. Urko M. Marigorta and Greg Gibson are supported by Project 3 of NIGMS P01 GM099568 (B. Weir, U. Washington, PI).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00225/abstract

REFERENCES

Abegunde, D. O., Mathers, C. D., Adam, T., Ortegon, M., and Strong, K. (2007). The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet* 370, 1929–1938. doi: 10.1016/S0140-6736(07)61696-1

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* 40, 616–622. doi: 10.1038/ng.109

Andreasen, C. H., Stender-Petersen, K. L., Mogensen, M. S., Torekov, S. S., Wegner, L., Andersen, G., et al. (2008). Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. *Diabetes* 57, 95–101. doi: 10.2337/db07-0910

Cutler, D. M., Glaeser, E. L., and Shapiro, J. M. (2003). Why have Americans become more obese? *J. Econ. Perspect.* 17, 93–118. doi: 10.1257/089533003769204371

Demerath, E. W., Lutsey, P. L., Monda, K. L., Linda Kao, W. H., Bressler, J., Pankow, J. S., et al. (2011). Interaction of FTO and physical activity level on adiposity in African-American and European-American adults: the ARIC study. *Obesity (Silver Spring)* 19, 1866–1872. doi: 10.1038/oby.2011.131

Do, R., Xie, C., Zhang, X., Mannisto, S., Harald, K., Islam, S., et al. (2011). The effect of chromosome 9p21 variants on cardiovascular disease may be modified by dietary intake: evidence from a case/control and a prospective study. *PLoS Med.* 8:e1001106. doi: 10.1371/journal.pmed.1001106

Franks, P. W., Pearson, E., and Florez, J. C. (2013). Gene-environment and gene-treatment interactions in type 2 diabetes: progress, pitfalls, and prospects. *Diabetes Care* 36, 1413–1421. doi: 10.2337/dc12-2211

Gibson, G. (2009). Decanalization and the origin of complex disease. *Nat. Rev. Genet.* 10, 134–140. doi: 10.1038/nrg2502

Gibson, G., and Dworkin, I. (2004). Uncovering cryptic genetic variation. *Nat. Rev. Genet.* 5, 681–690. doi: 10.1038/nrg1426

Heber, D., and Carpenter, C. L. (2011). Addictive genes and the relationship to obesity and inflammation. *Mol. Neurobiol.* 44, 160–165. doi: 10.1007/s12035-011-8180-6

Hermisson, J., and Wagner, G. P. (2004). The population genetic theory of hidden variation and genetic robustness. *Genetics* 168, 2271–2284. doi: 10.1534/genetics.104.029173

Hubacek, J. A., Pikhart, H., Peasey, A., Kubinova, R., and Bobak, M. (2011). FTO variant, energy intake, physical activity and basal metabolic rate in Caucasians. The HAPIEE study. *Physiol. Res.* 60, 175–183.

Kilpelainen, T. O., Qi, L., Brage, S., Sharp, S. J., Sonestedt, E., Demerath, E., et al. (2011). Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS Med.* 8:e1001116. doi: 10.1371/journal.pmed.1001116

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838. doi: 10.1038/nature09410

Luan, J. A., Wong, M. Y., Day, N. E., and Wareham, N. J. (2001). Sample size determination for studies of gene-environment interaction. *Int. J. Epidemiol.* 30, 1035–1040. doi: 10.1093/ije/30.5.1035

McGrath, J. J., Hannan, A. J., and Gibson, G. (2011). Decanalization, brain development and risk of schizophrenia. *Transl. Psychiatry* 1, e14. doi: 10.1038/tp.2011.16

Meyers, J. L., Cerda, M., Galea, S., Keyes, K. M., Aiello, A. E., Uddin, M., et al. (2013). Interaction between polygenic risk for cigarette use and environmental exposures in the Detroit Neighborhood Health Study. *Transl. Psychiatry* 3, e290. doi: 10.1038/tp.2013.63

Nettleton, J. A., Hivert, M. F., Lemaitre, R. N., McKeown, N. M., Mozaffarian, D., Tanaka, T., et al. (2013). Meta-analysis investigating associations between healthy diet and fasting glucose and insulin levels and modification by loci associated with glucose homeostasis in data from 15 cohorts. *Am. J. Epidemiol.* 177, 103–115. doi: 10.1093/aje/kws297

Ogden, C. L., Yanovski, S. Z., Carroll, M. D., and Flegal, K. M. (2007). The epidemiology of obesity. *Gastroenterology* 132, 2087–2102. doi: 10.1053/j.gastro.2007.03.052

Park, J. H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* 42, 570–575. doi: 10.1038/ng.610

Patel, C. J., Bhattacharya, J., and Butte, A. J. (2010). An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* 5:e10746. doi: 10.1371/journal.pone.0010746

Patel, C. J., Chen, R., Kodama, K., Ioannidis, J. P., and Butte, A. J. (2013). Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum. Genet.* 132, 495–508. doi: 10.1007/s00439-012-1258-z

Patel, C. J., and Ioannidis, J. P. (2014). Studying the elusive environment in large scale. *JAMA* 311, 2173–2174. doi: 10.1001/jama.2014.4129

Pontzer, H., Raichlen, D. A., Wood, B. M., Mabulla, A. Z., Racette, S. B., and Marlowe, F. W. (2012). Hunter-gatherer energetics and human obesity. *PLoS ONE* 7:e40503. doi: 10.1371/journal.pone.0040503

Qi, L., Cornelis, M. C., Zhang, C., Van Dam, R. M., and Hu, F. B. (2009). Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men. *Am. J. Clin. Nutr.* 89, 1453–1458. doi: 10.3945/ajcn.2008.27249

Qi, Q., Chu, A. Y., Kang, J. H., Huang, J., Rose, L. M., Jensen, M. K., et al. (2014). Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. *BMJ* 348:g1610. doi: 10.1136/bmj.g1610

Qi, Q., Chu, A. Y., Kang, J. H., Jensen, M. K., Curhan, G. C., Pasquale, L. R., et al. (2012). Sugar-sweetened beverages and genetic risk of obesity. *N. Engl. J. Med.* 367, 1387–1396. doi: 10.1056/NEJMoa,1203039

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rokholm, B., Silventoinen, K., Angquist, L., Skytthe, A., Kyvik, K. O., and Sorensen, T. I. (2011a). Increased genetic variance of BMI with a higher prevalence of obesity. *PLoS ONE* 6:e20816. doi: 10.1371/journal.pone.0020816

Rokholm, B., Silventoinen, K., Tynelius, P., Gamborg, M., Sorensen, T. I., and Rasmussen, F. (2011b). Increasing genetic variance of body mass index during the Swedish obesity epidemic. *PLoS ONE* 6:e27135. doi: 10.1371/journal.pone.0027135

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948. doi: 10.1038/ng.686

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* 44, 483–489. doi: 10.1038/ng.2232

Van Vliet-Ostaptchouk, J. V., Snieder, H., and Lagou, V. (2012). Gene-lifestyle interactions in obesity. *Curr. Nutr. Rep.* 1, 184–196. doi: 10.1007/s13668-012-0022-2

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24. doi: 10.1016/j.ajhg.2011.11.029

Wong, M. Y., Day, N. E., Luan, J. A., Chan, K. P., and Wareham, N. J. (2003). The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int. J. Epidemiol.* 32, 51–57. doi: 10.1093/ije/dyg002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 April 2014; paper pending published: 03 June 2014; accepted: 28 June 2014; published online: 21 July 2014.

*Citation: Marigorta UM and Gibson G (2014) A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. Front. Genet. **5**:225. doi: 10.3389/fgene.2014.00225*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 25 August 2014
doi: 10.3389/fgene.2014.00290

Disrupted human–pathogen co-evolution: a model for disease

Nuri Kodaman[1,2], Rafal S. Sobota[1,2], Robertino Mera[3], Barbara G. Schneider[3] and Scott M. Williams[1]*

[1] Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
[2] Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN, USA
[3] Division of Gastroenterology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Dana Crawford, Vanderbilt University, USA
Sebastien Gagneux, Swiss Tropical and Public Health Institute, Switzerland

***Correspondence:**
Scott M. Williams, Department of Genetics, Geisel School of Medicine, Dartmouth College, HB-6044, Hanover, NH 03755, USA
e-mail: scott.williams@dartmouth.edu

A major goal in infectious disease research is to identify the human and pathogenic genetic variants that explain differences in microbial pathogenesis. However, neither pathogenic strain nor human genetic variation in isolation has proven adequate to explain the heterogeneity of disease pathology. We suggest that disrupted co-evolution between a pathogen and its human host can explain variation in disease outcomes, and that genome-by-genome interactions should therefore be incorporated into genetic models of disease caused by infectious agents. Genetic epidemiological studies that fail to take both the pathogen and host into account can lead to false and misleading conclusions about disease etiology. We discuss our model in the context of three pathogens, *Helicobacter pylori*, *Mycobacterium tuberculosis* and human papillomavirus, and generalize the conditions under which it may be applicable.

Keywords: host–pathogen co-evolution, human disease, *Helicobacter pylori*, *Mycobacterium tuberculosis*, human papillomavirus, genome–genome interactions

INTRODUCTION

Human response to infectious agents is known to be highly heritable, but identifying the genetic variants responsible for differences in disease susceptibility has proven difficult. Pathogenic variation has, in some cases, become a better predictor of disease outcome, but it too does not sufficiently predict whether a given individual or class of individuals will present with disease. Thus far, genetic epidemiological studies of infectious disease have typically sought to explain the inter-individual variation in disease phenotypes by assessing genetic factors in humans or pathogens alone, under the implicit assumption that these factors have effects that are essentially independent of each other. Here, we argue that genome-by-genome interactions between host and pathogen are likely to play a major role in infectious disease etiology, and as such, should be incorporated into genetic epidemiological models. In short, insofar as host and pathogen jointly determine disease phenotypes, no genetic variant in either should be considered harmful without taking the context of the other into account.

The term "interaction" has two related but distinct meanings in the context of infectious disease, one molecular, and one statistical. Here we refer mainly to the statistical meaning of the term. At the individual level, all aspects of pathogenesis involve molecular interactions of varying importance, e.g., between a pathogenic epitope and a host receptor. Such interactions can be detected statistically, however, only when multiple variants exist in a population and when specific pairings lead to different effects. In some cases, pathogenic variants may function independently of host variation, and vice versa. However, because many pathogens have co-existed with their human hosts for millennia and have likely co-evolved with them, we argue here that statistical interactions, where appropriately sought, will often be found, with profound biomedical implications.

Recent advances in genomics have provided both the impetus and the means to evaluate human–pathogen co-evolutionary hypotheses directly. Whole-genome sequencing of many pathogenic species has substantially improved the resolution with which we classify strains, and facilitated the detection of potentially virulent genetic variants. A clearer picture of microbial evolution has also emerged, marked by selective mechanisms such as rapid gene gain/loss and horizontal gene transfer (Pallen and Wren, 2007). Overlaying human genetic variation onto this emerging evolutionary picture of microbial diversity offers the potential to make the pathogenic process more transparent.

The past few decades have also seen an explosion in studies seeking to identify human susceptibility loci for infectious diseases (Rowell et al., 2012). Candidate gene and family based linkage studies have identified several common polymorphisms with clinical significance at the population level, such as the *CCR5* deletion that protects against HIV (Samson et al., 1996; Picard et al., 2006; Casanova and Abel, 2007). However, most human susceptibility is in fact polygenic, with individual polymorphisms conferring small marginal effects (Hill, 2001). Where infectious disease phenotypes deviate from the "one susceptibility locus – one infection" model, elucidating the genetic architecture underlying inter-individual variation has proven elusive.

While genome-wide association studies (GWAS) may be better designed to accommodate multifactorial phenotypes, those performed thus far on infectious diseases have typically been less informative than GWAS performed on complex non-communicable diseases (Jallow et al., 2009; Hill, 2012; Ko and Urban, 2013). A major challenge facing the GWAS of infectious disease has been the recruitment of a sufficient number of cases and matched controls to achieve adequate statistical power (Hill, 2012; Ko and Urban, 2013). Another potential drawback, and the

one that concerns us here, is the fact that many infectious disease phenotypes depend on complex interactions between host and pathogen genomes. In such cases, the pooling together of human samples infected with even subtly different pathogenic strains can obscure genetic associations (Hill, 2012; Ko and Urban, 2013). A problem common to all GWAS is that the statistical effect sizes of biologically meaningful polymorphisms are often too small to pass significance thresholds after correction for multiple testing. This problem is exacerbated, however, when human polymorphisms (or networks of polymorphisms) (Wilfert and Schmid-Hempel, 2008) confer variable, or even opposite effects in the context of different pathogenic strains within the same study cohort. In this regard, it is perhaps telling that the most successful GWAS performed on infectious disease susceptibility to date have been on leprosy; the signal-to-noise ratios in these association studies may be higher because *Mycobacterium leprae* exhibits substantially less genetic heterogeneity than many other pathogens (Monot et al., 2009; Hill, 2012).

There is in fact strong empirical and theoretical justification for the hypothesis that the effects of susceptibility and virulence alleles in the respective gene pools of humans and pathogens are often contingent upon each other. The evolution of virulence is a dynamic process, easily perturbed by extrinsic variables over space and time, and therefore unlikely to follow the same trajectory in every population. For example, a spike in the density of hosts available for transmission can select for increased virulence, by reducing the cost of lethal harm (Anderson and May, 1982). If a pathogen is transmitted vertically (parent to child), the genetic factors that affect pathogenicity are "co-inherited" by host and pathogen, often promoting commensalism (Frank, 1996; Messenger et al., 1999). Even in these cases, the adventitious introduction of a microbial competitor can induce a commensal species to evolve a defensive toxin that harms the host, if only incidentally (Blaser and Kirschner, 2007; Frank and Schmid-Hempel, 2008). The evolution of defenses against pathogenic harm must also navigate fitness tradeoffs that vary with population, including tradeoffs pertaining to the correlated nature of complex traits (Lambrechts et al., 2006). As pathogens evolve rapidly, exerting strong selective pressures on different human populations, host phenotypes will respond in the *ad hoc* manner typical of evolution, limited by the available genetic variation at hand (Jacob, 1977). Whether the result is a steady-state equilibrium due to a perpetual "arms race" or a commensal detente, the same genes and pathways are unlikely to be involved in every population. As a consequence, when humans and pathogens migrate to new environments or admix, the ensuing disruption of co-evolutionary equilibria and loss of complementarity between host and pathogen genotypes may yield unpredictable and potentially deleterious biomedical consequences.

Our emphasis on the significance of mismatched traits is consistent with the genetic mosaic theory of co-evolution, which aims to account for why virtually all co-evolutionary interactions observed in natural populations show spatial variation in outcomes (Thompson et al., 2002; Thompson, 2014). The theory posits that co-evolution occurs in the context of geographically distinct "selection mosaics," each characterized by a unique genetic and environmental profile, where environmental variables can include both biotic and abiotic factors. Every selection mosaic progresses toward its own co-evolutionary equilibrium, while gene flow between selection mosaics ensures that patterns of maladaptation will be common and detectable where properly studied (Thompson et al., 2002; Ridenhour and Nuismer, 2007).

Despite the likely etiological importance of human–pathogen co-evolution, attempts at empirical confirmation have been rare. Indeed, "proof" of co-evolution poses a formidable challenge, requiring a demonstration of increased reproductive fitness in each species driven by reciprocal changes in two genomes over time (Woolhouse et al., 2002). Although these criteria have been met in laboratory studies and in some natural populations (Lenski and Levin, 1985; Little, 2002; Little et al., 2006), a similarly rigorous assessment of human–pathogen co-evolution must accommodate long generation times and the genetic and phenotypic complexity of the human traits under selection. Nonetheless, substantial phenomenological evidence consistent with human–pathogen co-evolution now exists, including evidence of spatial patterns of parallel genetic variation between species, and of correlated functional changes at the molecular level (Kraaijeveld et al., 1998; Lively and Dybdahl, 2000; Funk et al., 2000; Woolhouse et al., 2002). The collection of high-density genomic data in paired human–pathogen samples and improvements in phenotypic data, as well as advances in pathogen genomics, should soon enable more explicit tests of the concept.

Our aim here is to summarize the growing body of evidence in favor of the hypothesis that genetic interactions driven by host and pathogen co-evolution can have significant implications for genetic epidemiological studies and biomedicine. While this is not a novel hypothesis, it remains understudied. We also underscore how recent advances in genomic technology provide new opportunities to test for genome-by-genome interactions, and offer suggestions on how to incorporate them into more accurate genetic models of disease.

HELICOBACTER PYLORI

Studies of *Helicobacter pylori* provide perhaps the best evidence in favor of human–pathogen co-evolution, and distinctly illustrate the power of the modern genetic toolkit to investigate it. *H. pylori* chronically infects the gastric epithelia of half the world's population, causing peptic ulcers in 10–20% of those infected, and distal gastric carcinoma in ∼1% (Peek and Blaser, 2002; Jemal et al., 2011). The majority of individuals infected, however, suffer only from superficial gastritis in adulthood, while likely gaining protection against diseases such as esophageal cancer and reflux esophagitis, and more controversially, childhood asthma and diarrhea (Rothenbacher et al., 2000; Vaezi et al., 2000; Blaser et al., 2008). That *H. pylori* should have a largely innocuous and potentially symbiotic relationship with its host follows from co-evolutionary theory, based on its vertical mode of transmission, its long-term colonization of a single host, and its ∼50,000 year association with *Homo sapiens* (Rothenbacher et al., 2002; Moodley et al., 2012). Why a fraction of individuals develop life-threatening clinical disease, on the other hand, requires explanation, with one possibility being the disruption of long-standing co-evolutionary relationships.

Although *H. pylori*-mediated diseases often advance to the clinical stage in late adulthood, their onset typically occurs during reproductive years (Correa et al., 1976; Susser and Stein, 2002). Importantly, a disease need not have an especially large selection coefficient to shape allele frequency distributions in populations, especially over thousands of years (Ewald and Cochran, 2000). In fact, the historical fitness load of peptic ulcers, obtained by multiplying prevalence by selection coefficient, has been estimated to be similar to those for infectious diseases such as meningitis and rubella (Cochran et al., 2000). Also consistent with co-evolutionary theory is the fact that *H. pylori*-mediated gastric diseases occur disproportionately in men (Susser and Stein, 2002; Engel et al., 2003); *H. pylori* is usually, but not necessarily, transmitted by the mother, such that female fitness has likely exerted a stronger constraint against *H. pylori* virulence.

Some *H. pylori* virulence factors appear to increase the risk of serious clinical outcome regardless of host genotype. The *cag* pathogenicity island, present in some strains, encodes a type IV secretion system, and *VacA* encodes a pore-forming cytotoxin. Both have been implicated as carcinogenic risk factors, though neither is a necessary nor sufficient one (Wroblewski et al., 2010). Other virulence factors released by *H. pylori* include urease, which facilitates neutralization of the otherwise forbidding acidity of the gastric mucosa; NAP, which enables iron uptake; and arginase, which helps *H. pylori* subvert host macrophages. These, like most *H. pylori* virulence factors, operate to create a basal inflammatory state without generating an excessive immune response. Serious clinical disease reflects a disturbance of this balance (Baldari et al., 2005; Blaser and Kirschner, 2007; Salama et al., 2013).

The maintenance of this balance also depends partly on human genetic factors (Lichtenstein et al., 2000; Chiba et al., 2006; Mayerle et al., 2013a). Candidate gene studies on *H. pylori*-mediated diseases have implicated several gene polymorphisms that appear to affect risk, most notably in the interleukin-1 (IL-1) family of cytokines (Schneider et al., 2008). Recently, two GWAS assessing susceptibility to gastric cancer and *H. pylori* infection identified SNPs with odds ratios ranging from 1.3 to 1.4, mostly of uncertain biological function (Shi et al., 2011; El-Omar, 2013; Mayerle et al., 2013b, **Table 1**). These polymorphisms account for only a small proportion of the estimated heritability of disease phenotypes.

Studies of human or *H. pylori* genetics in isolation have generally failed to explain why populations with similar rates of *H. pylori* infection exhibit strikingly different susceptibilities to gastric cancer. For example, in many African and South Asian countries, the low incidences of gastric cancer in the presence of almost universal rates of *H. pylori* infection remain a source of much speculation, and have been referred to collectively as the "African enigma" and the "Asian enigma" (Holcombe, 1992; Campbell et al., 2001; Ghoshal et al., 2007). In Latin America, where *H. pylori* strains native to Amerindian populations have been largely displaced by European strains (Dominguez-Bello et al., 2008; Correa and Piazuelo, 2012), the predominantly Amerindian populations living at high altitudes suffer disproportionately from gastric cancer relative to other populations with similar infection rates (de Sablet et al., 2011; Torres et al., 2013). These and other points of evidence raise the possibility that the pathogenicity of a given *H. pylori* strain may vary with human genomic variation, and that some individuals may be better adapted to their infecting strains than others.

Modern genomic techniques have made the assessment of such hypotheses feasible. Over the past two decades, a comprehensive phylogeography of *H. pylori* has been constructed using multilocus sequence typing (MLST), a procedure by which polymorphisms in fragments from housekeeping genes are used to characterize bacterial isolates (Maiden et al., 1998). Analyses of samples from around the world have revealed a strong concordance between *H. pylori* phylogenetic clusters and the geographical locations from which they are derived (Falush et al., 2003; Moodley and Linz, 2009; Moodley et al., 2009). Ancestral *H. pylori* sequences inferred using MLST data also correspond to geographically defined human populations (Falush et al., 2003; Moodley et al., 2012). The typical modern *H. pylori* chromosome is now understood to be an amalgam of fragments from multiple ancestral sequences, a consequence of *H. pylori's* high recombinogenicity (Suerbaum et al., 1998; Falush et al., 2003). The genome of an *H. pylori* isolate can thus be quantitatively resolved into ancestral proportions, which correlate with proportions of human ancestry in admixed populations (Kodaman et al., 2014). In some cases, the ancestries of *H. pylori* isolates outperform human mitochondria in differentiating ethnic groups (Wirth et al., 2004).

These shared patterns of ancestry are unlikely to have arisen merely from parallel divergence due to founder effects or neutral drift. Certainly, the well-documented evolvability of functional loci within *H. pylori* strains, even within single individuals over a 6 year span, argues for the importance of adaptive microevolution (Israel et al., 2001; Dorer et al., 2009). Furthermore, at least 25% of known genes, including genes involved in mucosal adherence and the evasion of host immunity, are absent in some *H. pylori* strains isolated from different ethnic groups (Salama et al., 2000; Gressmann et al., 2005). In at least one case, variants of an *H. pylori* gene (*babA2*) encode adhesion proteins that exhibit host-specific effects, a hallmark of co-evolution. BabA binds to blood group antigens, triggering the release of proinflammatory cytokines. Notably, Amerindians, who almost all carry blood group O, harbor strains with a BabA variant that has up to a 1500-fold greater binding affinity to blood group O (Aspholm-Hurtig et al., 2004).

If we conclude from these patterns of genetic covariation that co-evolution between humans and *H. pylori* has occurred and that it has promoted commensalism, then we may ask whether individuals who develop serious clinical disease have inherited mutually ill-adapted sets of host and pathogen alleles. Under this hypothesis, we should expect to find significant interactions between specific pairs of host and pathogen loci in disease models. Toward this end, candidate pairs of loci can be tested based on biochemical evidence of protein–protein interactions, such as those between the adhesin BabA and the Lewis(b) antigen, its epithelial receptor (Backstrom et al., 2004). However, the effect size of any single two-locus interaction may be relatively small, as gastric disease etiology is phenotypically heterogeneous, and likely to be influenced by a large number of human and *H. pylori* genes (El-Omar, 2013). Thus, characterizing the relevant loci in a biologically meaningful way will ultimately require a systems biological approach.

Table 1 | Genetic variants identified by GWAS for phenotypes related to infection by *H. pylori, M. tuberculosis,* and human papillomavirus.

Disease/trait	Gene	SNP	Cases/controls	Population	*p*-value	OR[1]	95% CI[2]	Reference
Gastric cancer	*ZBTB20*	rs9841504	1006/2273	Chinese	1.7E-09	0.76	[0.69–0.83]	Shi et al. (2011)
Gastric cancer	*PRKAA1*	rs13361707	1006/2273	Chinese	7.6E-29	1.41	[1.32–1.49]	Shi et al. (2011)
H. pylori serologic status	*TLR10*	rs10004195	2623/7862	European	1.4E-18	0.70	[0.65–0.76]	Mayerle et al. (2013b)
H. pylori serologic status	*FCGR2A*	rs368433	2623/7862	European	2.1E-08	0.73	[0.65–0.85]	Mayerle et al. (2013b)
Tuberculosis	*RCN1–WT1*	rs2057178	2127/5636	African	2.6E-09	0.77	[0.71–0.84]	Thye et al. (2012)
Tuberculosis	*RPS4XP18–UBE2CP2*	rs4331426	2237/3122	African	6.8E-09	1.19	[1.13–1.27]	Thye et al. (2010)
Cervical cancer	*EXOC1*	rs13117307	1364/3028	Chinese	9.7E-09	1.26	[1.16–1.36]	Shi et al. (2013)
Cervical cancer	*HLA-DPB2*	rs4282438	1364/3028	Chinese	4.5E-27	0.75	[0.71–0.79]	Shi et al. (2013)
Cervical cancer	*ZPBP2–GSDMB*	rs8067378	1364/3028	Chinese	2.0E-08	1.18	[1.11–1.25]	Shi et al. (2013)
Cervical cancer	–	rs9277952	1364/3028	Chinese	2.3E-09	0.85	[0.81–0.90]	Shi et al. (2013)
Cervical cancer	*MICA*	rs2516448	2174/5002	European	1.6E-18	1.42	[1.31–1.54]	Chen et al. (2013)
Cervical cancer	*HLA-DRB1–HLA-DQA1*	rs9272143	2174/5006	European	9.3E-24	0.67	[0.62–0.72]	Chen et al. (2013)
Cervical cancer	*HLA-DPB2*	rs3117027	2171/4986	European	4.9E-08	1.25	[1.15–1.35]	Chen et al. (2013)

[1] *OR, odds ratio.*
[2] *CI, confidence interval.*

We recently took a broad-based view to assess the impact of human – *H. pylori* co-evolution on gastric disease, using ancestry estimates from both humans and their *H. pylori* isolates in the absence of knowledge of specific interacting loci (Kodaman et al., 2014). Our study participants were recruited from two Colombian populations with highly different rates of gastric cancer, despite a nearly universal prevalence of *H. pylori* infection in both. We found that the low-risk human, coastal population was of admixed African, European, and Amerindian ancestry, whereas the high-risk, Andean population was mainly of Amerindian ancestry, with a minority of European ancestry. Severity of gastric disease correlated with the proportion of African *H. pylori* ancestry in patients with primarily Amerindian ancestry. On the other hand, patients with a large proportion of African human ancestry infected by African *H. pylori* strains had the best prognoses, consistent with ancestral coadaptation, and likely pertinent to the "African enigma." The interaction between Amerindian human ancestry and African *H. pylori* ancestry accounted for the difference in disease risk between mountain and coastal populations, whereas even the well-known virulence factor, CagA, did not. These findings are thus consistent with the idea that neither human nor *H. pylori* genetic variation confers susceptibility or virulence *per se*, but only in context (**Figure 1**).

These findings also bring to light how understanding co-evolutionary interactions can inform and improve public health measures. It has been suggested that because *H. pylori* dominates the gastric microbiome in infected persons and has been shown to confer some beneficial effects, large-scale antibiotic eradication programs may not be warranted (Bik et al., 2006; Hung and Wong, 2009). Simply estimating ancestry from human samples and *H. pylori* isolates may help to identify individuals at greatest risk for gastric cancer, for whom antibiotic treatment may be most appropriate.

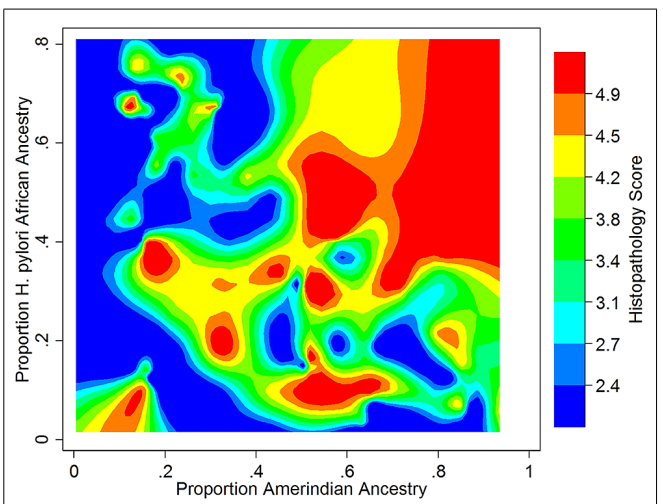


FIGURE 1 | Gastric histopathology as a function of Amerindian human and African *H. pylori* ancestry in a Colombian population (*N* = 121, age > 39). Histopathology was scored on a continuous scale, with 2 (blue) representing gastritis and 5 (red) representing dysplasia. Data from Kodaman et al. (2014). Reference samples from the 1000 Genomes Project (Abecasis et al., 2012), HapMap (The International HapMap Consortium, 2005), and the Human Genome Diversity Project (Cavalli-Sforza, 2005) were used to calculate human ancestry, and from the MLST database (Maiden et al., 1998) to calculate *H. pylori* ancestry.

MYCOBACTERIUM TUBERCULOSIS COMPLEX

Another interesting candidate to study from a co-evolutionary perspective is *Mycobacterium tuberculosis* (Mtb) and closely related

species, believed to have co-existed with anatomically modern humans for ∼70,000 years (Comas et al., 2013). Since the advent of antibiotics, tuberculosis (TB) has ceased to be as common a cause of human mortality as it once was, but it remains among the most deadly infectious diseases worldwide, with immunocompromised individuals at particularly high risk (Dye and Williams, 2010; Fenner et al., 2013). As with *H. pylori*, the majority of Mtb infections do not develop into clinical disease: 90% of cases are asymptomatic with only latent infection. However, 10% of individuals with latent infections develop TB over their lifetime, for mostly unknown reasons (Barry et al., 2009).

In contrast to *H. pylori,* Mtb is transmitted horizontally, and must cause active disease to be transmitted (e.g., *via* coughing or sneezing). Because Mtb transmission increases with virulence, evolutionary theory predicts that strong selective pressures should favor increased virulence until the number of transmissions per infected host reaches a fitness-reducing limit (Knolle, 1989; Frank and Schmid-Hempel, 2008). Such a limit necessarily depends on population-specific parameters, of which host density is probably the most important (Comas et al., 2013). Thus, the limited pathogenicity and chronicity of Mtb likely reflect its historical adaptation to isolated, low-density human populations. These historical conditions remain relevant in part because Mtb reproduces clonally and without lateral gene transfer; evolution only through point mutations and irreversible gene deletions limits a pathogen's ability to shift virulence strategies rapidly in response to changing population parameters (Achtman, 2008; Galagan, 2014).

Before advances in genotyping technology improved strain classification, the apparent genetic homogeneity of Mtb led investigators to believe that variation in disease outcome depended primarily on environmental and human genetic factors (Galagan, 2014). Twin and adoption studies provided compelling evidence for the involvement of human genetic variation as a risk modifier (Comstock, 1978). The most recent analyses have calculated the heritable component of Mtb-related immune response phenotypes to range from 30 to 71% (Moller and Hoal, 2010). These findings have motivated a large number of linkage and candidate gene association studies seeking to identify relevant susceptibility loci, but results have often been inconclusive or, worse, contradictory. Many biologically plausible genes, such as those that encode vitamin-D-binding protein (Lewis et al., 2005; Gao et al., 2010), the phagolysomal membrane protein NRAMP/SLC11A1 (Hoal et al., 2004; Velez et al., 2009), and the dendritic adhesion molecule DC-SIGN (Barreiro et al., 2006; Olesen et al., 2007), appear to associate with TB in some human populations, but not others. Inconsistent replication across ethnic groups has also beset the handful of GWAS performed on TB (Chimusa et al., 2014). The few loci that have passed genome-wide significance thresholds also lack clear biological interpretability and fail to explain more than a trivial portion of the estimated heritable component of TB susceptibility (Thye et al., 2010, 2012, **Table 1**).

Since the advent of PCR-based genotyping techniques, it has become increasingly clear that Mtb genetic variation is non-trivial and clinically consequential (Malik and Godfrey-Faussett, 2005; Nicol and Wilkinson, 2008). Most notably, strains now recognized as part of the "Beijing family," first genotyped in the 1990s following several drug-resistant outbreaks, have been found to exhibit greater efficiency of transmission and to cause more severe disease phenotypes in many animal models (Glynn et al., 2002; Reed et al., 2004; Parwati et al., 2010). Whole-genome sequencing of a large number of clinical Mtb isolates has since revealed over 30,000 Mtb SNPs, a large proportion of which are non-synonymous (Comas et al., 2013; Stucki and Gagneux, 2013). It has been shown that even a few such SNPs can shift a strain from avirulent to virulent (Reiling et al., 2013).

High-throughput sequence data have also enabled the construction of a robust phylogenetic tree, the major branches of which parallel human mitochondrial phylogeny (Comas et al., 2013). Seven major human-adapted Mtb lineages have now been identified, which can be classified as "ancient" or "modern" (Hershberg et al., 2008; Comas et al., 2013). The Beijing family of strains, which causes 50% of infections in East Asia and 13% worldwide, belongs to the most modern lineage. In contrast, *Mycobacterium africanum*, which causes up to half of TB cases in West Africa, belongs to the most ancient Mtb clade, its divergence predating the human migration out of Africa (de Jong et al., 2010). Although strains within all major Mtb lineages induce an overlapping range of immune responses, clade-specific patterns of virulence are emerging. For example, evolutionarily modern lineages appear to induce a less severe early inflammatory response, which possibly increases the efficiency of transmission (Moller and Hoal, 2010; Portevin et al., 2011). A large number of studies in experimental models have also confirmed that diverse Mtb strains reflect substantial functional diversity (Coscolla and Gagneux, 2010).

It is thus likely that genetic factors in both Mtb and humans influence a wide range of TB phenotypes, including those pertaining to infectivity, progression from latent to active disease, and effectiveness of treatment (de Jong et al., 2008; Comas and Gagneux, 2011). However, whether Mtb genetic variation influences disease outcome independently of human genetic variation, and vice versa, is a question that has only recently been addressed (Gagneux, 2012). The mirrored pattern of human and Mtb phylogeography indicates that co-evolution has likely occurred, and consequently, that genome-by-genome interactions may be significant. However, identifying these interactions and assessing their clinical relevance requires the demonstration of heterogeneous outcomes in paired human and Mtb samples of multiple genotypic backgrounds. A small number of published studies to date have met this criterion, assessing previously implicated loci (e.g., in immunogenicity pathways). A study in a Vietnamese cohort found that a variant of the Toll-interleukin 2 receptor (TLR2), known to trigger a cytokine cascade upon recognition of Mtb, increased TB susceptibility only in patients infected with a Beijing strain (Caws et al., 2008). In a Ghanaian cohort, a polymorphism in the immunity-related GTPase M (*IRGM*) gene conferred protection against the European lineage of *M. tuberculosis*, but not *M. africanum* (Intemann et al., 2009). Perhaps of consequence, a gene deletion in the European Mtb strains increases their vulnerability to the autophagy pathway, mediated by IRGM. Thus, the high frequency of the human *IRGM* polymorphism in West Africa has been proposed to explain the competitive advantage of *M. africanum* there (Intemann et al., 2009). The innate immunity-related genes *ALOX5* and *MBL* have also been shown to influence

the infectivity of *M. africanum*, but not other strains, in Ghanaian populations (Herb et al., 2008; Thye et al., 2011).

Despite being an ancient strain with ample opportunity to spread beyond West Africa, *M. africanum* has not done so, possibly indicating host-specific adaptation (de Jong et al., 2010; Gagneux, 2012). Other Mtb lineages also appear to associate preferentially with particular human populations, though not as exclusively. A study of ethnically diverse, US-born patients in San Francisco showed that such preferential associations with Mtb lineages persisted even in a cosmopolitan setting (Gagneux et al., 2006). Interestingly, when TB transmission in non-sympatric populations did occur, patients were significantly more likely to be immunocompromised, indicating that non-sympatric Mtb lineages may require some degree of host immunosuppression to compete with sympatric lineages. Mechanisms of Mtb immune evasion, therefore, may have been shaped by population-specific variation in human immune response.

While the above discussion has focused mainly on pulmonary TB, we note here that extra-pulmonary TB, a less common and more severe form of disease, may be especially amenable to analyses guided by co-evolutionary hypotheses. This form of the disease leads more quickly to fatality and results in fewer transmissions than the pulmonary form (Sharma and Mohan, 2004), which probably represents a non-optimal outcome in terms of Mtb fitness. However, data on extra-pulmonary TB to support co-evolutionary hypotheses – especially historical data pre-dating the antibiotic era and the HIV epidemic – are at present lacking (Tiemersma et al., 2011).

HUMAN PAPILLOMAVIRUS

Human papillomavirus (HPV) is the most common sexually transmitted infectious agent in the world, and the second most common infectious cause of cancer after *H. pylori* (de Martel et al., 2012). Cervical cancer is the major source of mortality associated with HPV, but the virus also causes cancers of the anus, vagina, penis, and oropharynx (zur Hausen, 1989; zur Hausen, 1991; Carter et al., 2001; de Martel et al., 2012). Although over 100 types of papillomaviruses infect humans, only a fraction of them are carcinogenic (Bernard et al., 2010). Infection with two specific types, HPV 16 and HPV 18, account for approximately 70% of cervical cancer cases worldwide, with the remainder of cases largely attributable to 14 other types (Bernard et al., 2010). Nevertheless, the great majority of infections with even carcinogenic HPV types are ultimately benign, demonstrating that HPV infection, although necessary, is not sufficient to cause of cervical cancer (Schiffman et al., 2005; Plummer et al., 2007).

Papillomaviruses (PVs) are notable for their slow rate of evolution relative to other pathogens – only an order of magnitude higher than humans, in the case of HPV (Ong et al., 1993; Rector et al., 2007; Shah et al., 2010). This is commonly attributed to their use of high-fidelity host replication mechanisms (Van Doorslaer, 2013). A slow evolutionary rate precludes rapid adaptation to new hosts, and PV strains correspondingly show little evidence of interspecies transmission or related horizontal gene transfer (Herbst et al., 2009; Shah et al., 2010; Van Doorslaer, 2013). All carcinogenic types of HPV belong to a single genus of papillomaviruses that diverged from a common ancestor about 75 million years ago, predating the primate lineage (Rector et al., 2007; Van Doorslaer, 2013). By the emergence of *H. sapiens*, the common ancestor of HPV 16 and HPV 18 had diverged into separate species, and in fact HPV 16 and HPV 18 had already diverged from all other HPV types within their respective species clades (Lewin, 1993; Ong et al., 1993). Given this combination of early divergence, slow evolution, and strict host specialization, we would expect variants within HPV types independently to have similar phylogeographic patterns to that of *H. sapiens*. Global data collected for the two most frequently sexually transmitted types, HPV 16 and 18, reflect such a pattern (Bernard, 1994). The subtypes and variants of HPV 16 cluster into five major branches of a phylogenetic tree: European (E), Asian/American (AA), East Asian (As), and two African (Af1 and Af2) (Ho et al., 1993; Ong et al., 1993). Subtypes and variants of HPV-18 clustering into three major branches: African (Af), European (E), and Asian + American Indian (As+AI) (Ong et al., 1993).

Biochemical and bioinformatic analyses indicate that HPV evolution has not been entirely neutral. Viral genes expressed early during a PV infection, for example, appear to have evolved at different rates than those expressed late (Garcia-Vallve et al., 2005; Rector et al., 2007). Although most PV genes show signs of strong purifying selection, the exceptions appear to be important (DeFilippis et al., 2002; Chen et al., 2005; Carvajal-Rodriguez, 2008). Two genes under diversifying selection, *E6* and *E7*, are essential for viral replication. They induce cell cycle progression in host cells, and encode proteins that, in the high-risk HPVs, are oncogenic (White et al., 1994; Doorbar, 2006; Klingelhutz and Roman, 2012). Of note, E6 and E7 interfere with the human tumor suppressor proteins, pRB and p53 (Dyson et al., 1989; Huibregtse et al., 1993a,b; Storey et al., 1998; Munger et al., 2004; Doorbar, 2006). In turn, polymorphisms in the human p53 gene were shown to modulate the tumorigenicity of HPV 16 and 18 (Storey et al., 1998). Patients homozygous for the p53Arg mutation were seven times more likely to develop cervical cancer than individuals with 1 or 2 p53Pro alleles (Storey et al., 1998). Other human polymorphisms, such as those in the genes *RPS* and *TYMS*, influence HPV transmissibility. In a study of high-risk HPV infections in Nigerian women, variants in these genes were shown to modulate risk of infection with HPV 16 and 18. Despite the effects described above, genetic variation in neither the host nor the pathogen has been successful in explaining most heritable risk of HPV-associated disease, when considered in isolation (Magnusson et al., 2000; Hildesheim and Wang, 2002; Wheeler, 2008; Chen et al., 2013; Shi et al., 2013, **Table 1**).

Because the integration of the HPV genome within the human genome is permanent, death of the host ends all possibility of viral multiplication and transmission. Even strains that damage the health of the host sufficiently to reduce human-to-human sexual contact can suffer a competitive disadvantage. Therefore, both host and pathogen should cooperate to prevent severe disease. As with *H. pylori* and MTB, there is some empirical evidence supporting the idea that humans and HPV types co-evolved to limit tumorigenesis, and that evolutionarily mismatched strains may be driving severe clinical outcomes. A study of high-grade

cervical intraepithelial neoplasia (CIN) and invasive cervical cancer in an Italian cohort of Caucasian women demonstrated that non-European variants of HPV16, Af1 and AA, were found at an increased frequency in invasive lesions (Tornesello et al., 2004). A separate study of mostly Caucasian (81%) female university students in the United States showed that those infected with non-European HPV 16 variants were 6.5 times more likely to develop high-grade CIN than those with European variants (Xi et al., 1997). The same study demonstrated a similar HPV 16-related risk profile (4.5 relative risk) in a predominantly Caucasian (79%) population of women presenting at a sexually transmitted disease clinic (Xi et al., 1997). Finally, at the molecular level, there is some evidence that variants of the HPV 16 E6 protein, described above, may be better adapted for replication within specific hosts (DeFilippis et al., 2002).

DISCUSSION

Taken together, the three examples above illustrate how co-evolution can promote a reduction in antagonism between pathogen and host, and in doing so leave discernible signatures on the genomes of both species. If, as we argue here, the disruption of historical co-evolutionary relationships can explain many differences in disease outcomes, knowledge of the conditions under which such relationships arise and dissolve will be helpful in defining genetic architecture of disease etiology. The applicability of this model depends, to a large extent, on the degree of integration between host and pathogen genomes, which can take many forms.

A long-standing association between humans and pathogens may be a necessary factor for cross-genomic integration, as with the three pathogens we have discussed. In contrast, many infectious diseases that occur epidemically are caused by zoonotic pathogens for which the human host is an evolutionary dead end, such as *Salmonella enterica* and *Borrelia burgdorferi* (Sokurenko et al., 2006; Falush, 2009). Other pathogens have had limited occasion to co-evolve with humans, because they cause disease primarily on an opportunistic basis (e.g., *Streptococcus pneumonia* or *Clostridium difficile*) or over a broad range of hosts (e.g., *Toxoplasma gondii*) (Ajzenberg et al., 2004; Sokurenko et al., 2006). The epidemic outbreaks caused by these pathogens may leave detectable signatures on the human genome, but reciprocal evolution in the pathogen need not occur.

For human-specific pathogens that cause endemic diseases and are not recent, the likelihood that severe disease is the outcome of a co-evolutionary mismatch should increase with the overlap between host and pathogen fitness. The pathogenicity of vertically transmitted pathogens, for example, should decrease over time, because such pathogens often depend on host survival (and possibly reproduction) for transmission. However, a strong overlap between host and pathogen fitness can also exist in the absence of vertical transmission. A horizontally transmitted pathogen, such as HPV, can evolve to be largely benign insofar as it depends on a healthy host for transmission.

When a pathogen's fitness depends on its ability to cause damage to its human host, as with Mtb, attenuated antagonism becomes a special case, and its disruption becomes more difficult to detect and requires more evidence to confirm. While Mtb strains that increase the duration of a transmissible state will generally have a competitive advantage, the optimal duration can be expected to vary based on many population-level parameters, such as host density. This probably explains why modern Mtb lineages that are more common in high-density urban populations exhibit greater virulence. On the other hand, if horizontal transfer is confined to small, isolated populations, it may be considered effectively vertical. With such pathogens, a better understanding of the co-evolutionary history will be necessary to infer whether severe disease is caused by disrupted co-evolution or by another factor, such as infection by a universally more virulent strain or an opportunistic infection in an immunosuppressed patient.

The life history of the pathogen is also important in assessing the possibility and nature of co-evolution. A pathogen typically faces a tradeoff between fecundity and longevity. Increased fecundity within a host increases the probability (or rate) of transmission, but may negatively affect host lifespan or mobility (Frank and Schmid-Hempel, 2008). Therefore, a pathogen's position on the continuum between greater fecundity and increased longevity will often reflect the degree to which its fitness depends on the health of the host. The case of HPV is somewhat of an exception in this regard. Host immune responses can induce diverse strategies, creating HPV types that are highly fecund, or less fecund with few virions per host. Whereas highly fecund types are more likely to transmit, they are also more likely to induce a vigorous immune response leading to clearance. Low fecundity types on the other hand, are more likely to persist as subclinical infections that can lead to prolonged inflammation and eventually cancer (DeFilippis et al., 2002). However, human populations that co-evolved with specific variants of these persistent types may be less likely to develop cancer, as described above.

Another factor influencing the applicability of the model we propose is a pathogen's recombinogenicity. In theory, a pathogen that recombines freely is more likely to be panmictic, and hence less likely to co-evolve with a particular human host population (Bull et al., 1991). In fact, epidemic disease outbreaks often follow recombination events, and the pathogens responsible for the epidemics often appear superficially clonal, likely reflecting the rapid proliferation of especially successful recombinant strains (Grigg et al., 2001; Heitman, 2006). A case in point is *Neisseria meningitides* (Falush, 2009), as well as the eukaryotic parasites *Toxoplasma gondii* and *Plasmodium falciparum*, which though able to recombine sexually, exhibit surprisingly limited genetic diversity (Grigg et al., 2001). On the other hand, the strict clonality of Mtb and HPV has likely favored co-evolution, leading to reduced antagonism, while recombination in *H. pylori* can disrupt the co-evolutionary relationship favored by vertical transmission.

Recombination can also occur via horizontal gene transfer, as among species within the microbiome (Smillie et al., 2011; Ravel et al., 2011; Liu et al., 2012). This would suggest that co-evolution might be a relatively weak force in shaping microbiotal genetic variation. However, data possibly supporting human–microbiome co-evolution exist; for example, the strongest correlate of an individual's microbiotal identity is ethnicity (Benson et al., 2010; Human Microbiome Project Consortium, 2012). The extent to which this correlation is driven by mutual genetic factors is unclear,

as recurring environmental exposure and frequent vertical transmission may also account for most, if not all of it (Turnbaugh et al., 2009). Assessing whether the genomes of the microbiome and humans are integrated will be a key area of research, as it relates to co-evolution and disease risk (McFall-Ngai et al., 2013).

CONCLUSION

While the prospect of introducing co-evolutionary interactions into genetic epidemiology models may appear to add a new layer of complexity to an already difficult problem, a co-evolutionary perspective should help us construct more precise and accurate hypotheses, improving our ability to find real and reproducible results. Importantly, co-evolved genes will not be neutral in either species, which may make their identification easier. Although many methods exist to find loci that are candidates to have evolved under selection (Aguileta et al., 2009; Karlsson et al., 2014), and these methods can assess the strength, timing, and direction of selection (e.g., balancing or positive), they are not at present well adapted to the study of joint patterns of selection.

If the ultimate goal is to find interacting genes that have co-evolved to be benign and are subsequently disrupted in disease, we will need to identify differential patterns of concerted selection in paired human and pathogenic loci from different populations. The limiting factor to the development of appropriate methods toward this end has probably been the lack of prospectively collected paired genetic data for humans and pathogens. Once these data are available, existing methods to detect epistasis within a species can be adapted for cross-species analyses in the absence of *a priori* biological hypotheses. Where evidence for selection exists, genetic variants can be filtered prior to analyses to detect epistasis. Framing hypotheses in the context of biochemical and bioinformatic functional evidence or pre-existing evidence for association can hone study design even further. For example, using paired data and pathogenic genetic variation as the outcome variable, novel epitopes have been discovered in association studies (Bartha et al., 2013). Such data can be used to mitigate the immense multiple testing burden incurred by a hypothesis-free approach to detecting genetic interactions.

Finally, we should note that the ultimate impact of this approach may extend beyond infectious diseases to what are traditionally considered non-communicable diseases. For example, we now recognize that both gastric and cervical cancers, as well as atherosclerosis, may have origins in infection (Libby et al., 2002; Porta et al., 2011). The number of such examples will certainly expand.

ACKNOWLEDGMENTS

This study was supported by the National Center for Research Resources, grant UL1 RR024975-01, which is now at the National Center for Advancing Translational Sciences; National Cancer Institute Grant P01 CA28842, the Vanderbilt-Ingram Cancer Center, the Wendy Dio family and the TJ. Martell Foundation. Scott M. Williams was partially supported by P20 GM103534.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* 62, 53–70. doi: 10.1146/annurev.micro.62.081307.162832

Aguileta, G., Refregier, G., Yockteng, R., Fournier, E., and Giraud, T. (2009). Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect. Genet. Evol.* 9, 656–670. doi: 10.1016/j.meegid.2009.03.010

Ajzenberg, D., Banuls, A. L., Su, C., Dumetre, A., Demar, M., Carme, B., et al. (2004). Genetic diversity, clonality and sexuality in *Toxoplasma gondii*. *Int. J. Parasitol.* 34, 1185–1196. doi: 10.1016/j.ijpara.2004.06.007

Anderson, R. M., and May, R. M. (1982). Coevolution of hosts and parasites. *Parasitology* 85, 411–426. doi: 10.1017/S0031182000055360

Aspholm-Hurtig, M., Dailide, G., Lahmann, M., Kalia, A., Ilver, D., Vikström, S., et al. (2004). Functional adaptation of BabA, the *H. pylori* ABO blood group antigen binding adhesin. *Science* 305, 519–522. doi: 10.1126/science.1098801

Backstrom, A., Lundberg, C., Kersulyte, D., Berg, D. E., Boren, T., and Amqvist, A. (2004). Metastability of *Helicobacter pylori* bab adhesin genes and dynamics in Lewis b antigen binding. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16923–16928. doi: 10.1073/pnas.0404817101

Baldari, C. T., Lanzavecchia, A., and Telford, J. L. (2005). Immune subversion by *Helicobacter pylori*. *Trends Immunol.* 26, 199–207. doi: 10.1016/j.it.2005.01.007

Barreiro, L. B., Neyrolles, O., Babb, C. L., Tailleux, L., Quach, H., McElreavey, K., et al. (2006). Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis. *PLoS Med.* 3:e20. doi: 10.1371/journal.pmed.0030020

Barry, C. E. III, Boshoff, H. I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., et al. (2009). The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat. Rev. Microbiol.* 7, 845–855. doi: 10.1038/nrmicro2236

Bartha, I., Carlson, J. M., Brumme, C. J., McLaren, P. J., Brumme. Z. L., John, M., et al. (2013). A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife (Cambridge)* 2:e01123. doi: 10.7554/eLife.01123

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

Bernard, H. U. (1994). Coevolution of papillomaviruses with human populations. *Trends Microbiol.* 2, 140–143. doi: 10.1016/0966-842X(94)90602-5

Bernard, H. U., Burk, R. D., Chen, Z., van Doorslaer, K., zur Hausen, H., and de Villiers, E. M. (2010). Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 401, 70–79. doi: 10.1016/j.virol.2010.02.002

Bik, E. M., Eckburg, P. B., Gill, S. R., Nelson, K. E., Purdom, E. A., Francois, F., et al. (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl. Acad. Sci. U.S.A.* 103, 732–737. doi: 10.1073/pnas.0506655103

Blaser, M. J., and Kirschner, D. (2007). The equilibria that allow bacterial persistence in human hosts. *Nature* 449, 843–849. doi: 10.1038/nature06198

Blaser, M. J., Chen, Y., and Reibman, J. (2008). Does *Helicobacter pylori* protect against asthma and allergy? *Gut* 57, 561–567. doi: 10.1136/gut.2007.133462

Bull, J. J., Molineux, I. J., and Rice, W. R. (1991). Selection of benevolence in a host–parasite system. *Evolution* 45, 875–882. doi: 10.2307/2409695

Campbell, D. I., Warren, B. F., Thomas, J. E., Figura, N., Telford, J. L., and Sullivan, P. B. (2001). The African enigma: low prevalence of gastric atrophy, high prevalence of chronic inflammation in West African adults and children. *Helicobacter* 6, 263–267. doi: 10.1046/j.1083-4389.2001.00047.x

Carter, J. J., Madeleine, M. M., Shera, K., Schwartz, S. M., Cushing-Haugen, K. L., Wipf, G. C., et al. (2001). Human papillomavirus 16 and 18 L1 serology compared across anogenital cancer sites. *Cancer Res.* 61, 1934–1940.

Carvajal-Rodriguez, A. (2008). Detecting recombination and diversifying selection in human alpha-papillomavirus. *Infect Genet. Evol.* 8, 689–692. doi: 10.1016/j.meegid.2008.07.002

Casanova, J. L., and Abel, L. (2007). Human genetics of infectious diseases: a unified theory. *EMBO J.* 26, 915–922. doi: 10.1038/sj.emboj.7601558

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* 6, 333–340. doi: 10.1038/nrg1596

Caws, M., Thwaites, G., Dunstan, S., Hawn, T. R., Lan, N. T., Thuong, N. T., et al. (2008). The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. *PLoS Pathog.* 4:e1000034. doi: 10.1371/journal.ppat.1000034

Chen, D., Juko–Pecirep, I., Hammer, J., Ivansson, E., Enroth, S., Gustavsson, I., et al. (2013). Genome-wide association study of susceptibility loci for cervical cancer. *J. Natl. Cancer Inst.* 105, 624–633. doi: 10.1093/jnci/djt051

Chen, Z., Terai, M., Fu, L., Herrero, R., DeSalle, R., and Burk, R. D. (2005). Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. *J. Virol.* 79, 7014–7023. doi: 10.1128/JVI.79.11.7014-7023.2005

Chiba, T., Seno, H., Marusawa, H., Wakatsuki, Y., and Okazaki, K. (2006). Host factors are important in determining clinical outcomes of *Helicobacter pylori* infection. *J. Gastroenterol.* 41, 1–9. doi: 10.1007/s00535-005-1743-4

Chimusa, E. R., Zaitlen, N., Daya, M., Moller, M., van Helden, P. D., Mulder, N. J., et al. (2014). Genome-wide association study of ancestry-specific TB risk in the South African coloured population. *Hum. Mol. Genet.* 23, 796–809. doi: 10.1093/hmg/ddt462

Cochran, G. M., Ewald, P. W., and Cochran, K. D. (2000). Infectious causation of disease: an evolutionary perspective. *Perspect. Biol. Med.* 43, 406–448. doi: 10.1353/pbm.2000.0016

Comas, I., and Gagneux, S. (2011). A role for systems epidemiology in tuberculosis research. *Trends Microbiol.* 19, 492–500. doi: 10.1016/j.tim.2011.07.002

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., et al. (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* 45, 1176–1182. doi: 10.1038/ng.2744

Comstock, G. W. (1978). Tuberculosis in twins: a re-analysis of the Prophit survey. *Am. Rev. Respir. Dis.* 117, 621–624.

Correa, P., Cuello, C., Duque, E., Burbano, L. C., Garcia, F. T., Botanos, O., et al. (1976). Gastric cancer in Colombia. III. Natural history of precursor lesions. *J. Natl. Cancer Inst.* 57, 1027–1035. doi: 10.1093/jnci/57.5.1027

Correa, P., and Piazuelo, M. B. (2012). Evolutionary history of the *Helicobacter pylori* genome: implications for gastric carcinogenesis. *Gut Liver* 6, 21–28. doi: 10.5009/gnl.2012.6.1.21

Coscolla, M., and Gagneux, S. (2010). Does *M. tuberculosis* genomic diversity explain disease diversity? *Drug Discov. Today Dis. Mech.* 7:e43–e59. doi: 10.1016/j.ddmec.2010.09.004

DeFilippis, V. R., Ayala, F. J., and Villarreal, L. P. (2002). Evidence of diversifying selection in human papillomavirus type 16 E6 but not E7 oncogenes. *J. Mol. Evol.* 55, 491–499. doi: 10.1007/s00239-002-2344-y

de Jong, B. C., Antonio, M., and Gagneux, S. (2010). *Mycobacterium africanum* – review of an important cause of human tuberculosis in West Africa. *PLoS Negl. Trop. Dis.* 4:e744. doi: 10.1371/journal.pntd.0000744

de Jong, B. C., Hill, P. C., Aiken, A., Awine, T., Antonio, M., Adetifa, I. M., et al. (2008). Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J. Infect. Dis.* 198, 1037–1043. doi: 10.1086/591504

de Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D., et al. (2012). Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.* 13, 607–615. doi: 10.1016/S1470-2045(12)70137-7

de Sablet, T., Piazuelo, M. B., Shaffer, C. L., Schneider, B. G., Asim, M., Chaturvedi, R., et al. (2011). Phylogeographic origin of *Helicobacter pylori* is a determinant of gastric cancer risk. *Gut* 60, 1189–1195. doi: 10.1136/gut.2010.234468

Dominguez-Bello, M. G., Perez, M. E., Bortolini, M. C., Salzano, F. M., Pericchi, L. R., Zambrano-Guzmán, O., et al. (2008). Amerindian *Helicobacter pylori* strains go extinct, as european strains expand their host range. *PLoS ONE* 3:e3307. doi: 10.1371/journal.pone.0003307

Doorbar, J. (2006). Molecular biology of human papillomavirus infection and cervical cancer. *Clin. Sci. (Lond.)* 110, 525–541. doi: 10.1042/CS20050369

Dorer, M. S., Talarico, S., and Salama, N. R. (2009). *Helicobacter pylori*'s unconventional role in health and disease. *PLoS Pathog.* 5:e1000544. doi: 10.1371/journal.ppat.1000544

Dye, C., and Williams, B. G. (2010). The population dynamics and control of tuberculosis. *Science* 328, 856–861. doi: 10.1126/science.1185449

Dyson, N., Howley, P. M., Munger, K., and Harlow, E. (1989). The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* 243, 934–937. doi: 10.1126/science.2537532

El-Omar, E. M. (2013). *Helicobacter pylori* susceptibility in the GWAS era. *JAMA* 309, 1939–1940. doi: 10.1001/jama.2013.5590

Engel, L. S., Chow, W. H., Vaughan, T. L., Gammon, M. D., Risch, H. A., Stanford, J. L., et al. (2003). Population attributable risks of esophageal and gastric cancers. *J. Natl. Cancer Inst.* 95, 1404–1413. doi: 10.1093/jnci/djg047

Ewald, P. W., and Cochran, G. M. (2000). *Chlamydia pneumoniae* and cardiovascular disease: an evolutionary perspective on infectious causation and antibiotic treatment. *J. Infect. Dis.* 181(Suppl. 3), S394–S401. doi: 10.1086/315602

Falush, D. (2009). Toward the use of genomics to study microevolutionary change in bacteria. *PLoS Genet.* 5:e1000627. doi: 10.1371/journal.pgen.1000627

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., et al. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science* 299, 1582–1585. doi: 10.1126/science.1080857

Fenner, L., Egger, M., Bodmer, T., Furrer, H., Ballif, M., Battegay, M., et al. (2013). HIV infection disrupts the sympatric host–pathogen relationship in human tuberculosis. *PLoS Genet.* 9:e1003318. doi: 10.1371/journal.pgen.1003318

Frank, S. A. (1996). Models of parasite virulence. *Q. Rev. Biol.* 71, 37–78. doi: 10.1086/419267

Frank, S. A., and Schmid-Hempel, P. (2008). Mechanisms of pathogenesis and the evolution of parasite virulence. *J. Evol. Biol.* 21, 396–404. doi: 10.1111/j.1420-9101.2007.01480.x

Funk, D. J., Helbling, L., Wernegreen, J. J., and Moran, N. A. (2000). Intraspecific phylogenetic congruence among multiple symbiont genomes. *Proc. Biol. Sci.* 267, 2517–2521. doi: 10.1098/rspb.2000.1314

Gagneux, S. (2012). Host–pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 850–859. doi: 10.1098/rstb.2011.0316

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S., et al. (2006). Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2869–2873. doi: 10.1073/pnas.0511240103

Galagan, J. E. (2014). Genomic insights into tuberculosis. *Nat. Rev. Genet.* 15, 307–320. doi: 10.1038/nrg3664

Gao, L., Tao, Y., Zhang, L., and Jin, Q. (2010). Vitamin D receptor genetic polymorphisms and tuberculosis: updated systematic review and meta-analysis. *Int. J. Tuberc. Lung Dis.* 14, 15–23.

Garcia-Vallve, S., Alonso, A., and Bravo, I. G. (2005). Papillomaviruses: different genes have different histories. *Trends Microbiol.* 13, 514–521. doi: 10.1016/j.tim.2005.09.003

Ghoshal, U. C., Tripathi, S., and Ghoshal, U. (2007). The Indian enigma of frequent *H. pylori* infection but infrequent gastric cancer: is the magic key in Indian diet, host's genetic make up, or friendly bug? *Am. J. Gastroenterol.* 102, 2113–2114. doi: 10.1111/j.1572-0241.2007.01324_13.x

Glynn, J. R., Whiteley, J., Bifani, P. J., Kremer, K., and van Soolingen, D. (2002). Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg. Infect. Dis.* 8, 843–849. doi: 10.3201/eid0805.020002

Gressmann, H., Linz, B., Ghai, R., Pleissner, K. P., Schlapbach, R., Yamaoka, Y., et al. (2005). Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS Genet.* 1:e43. doi: 10.1371/journal.pgen.0010043

Grigg, M. E., Bonnefoy, S., Hehl, A. B., Suzuki, Y., and Boothroyd, J. C. (2001). Success and virulence in toxoplasma as the result of sexual recombination between two distinct ancestries. *Science* 294, 161–165. doi: 10.1126/science.1061888

Heitman, J. (2006). Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* 16, R711–R725. doi: 10.1016/j.cub.2006.07.064

Herb, F., Thye, T., Niemann, S., Browne, E. N., Chinbuah. M. A., Gyapong, J., et al. (2008). ALOX5 variants associated with susceptibility to human pulmonary tuberculosis. *Hum. Mol. Genet.* 17, 1052–1060. doi: 10.1093/hmg/ddm378

Herbst, L. H., Lenz, J., Van Doorslaer, K., Chen, Z., Stacy, B. A., Wellenhan, J. F. Jr., et al. (2009). Genomic characterization of two novel reptilian papillomaviruses, Chelonia mydas papillomavirus 1 and Caretta caretta papillomavirus 1. *Virology* 383, 131–135. doi: 10.1016/j.virol.2008.09.022

Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., et al. (2008). High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* 6:e311. doi: 10.1371/journal.pbio.0060311

Hildesheim, A., and Wang, S. S. (2002). Host and viral genetics and risk of cervical cancer: a review. *Virus Res.* 89, 229–240. doi: 10.1016/S0168-1702(02)00191-0

Hill, A. V. (2001). The genomics and genetics of human infectious disease susceptibility. *Annu. Rev. Genomics Hum. Genet.* 2, 373–400. doi: 10.1146/annurev.genom.2.1.373

Hill, A. V. (2012). Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 840–849. doi: 10.1098/rstb.2011.0275

Hoal, E. G., Lewis, L. A., Jamieson, S. E., Tanzer, F., Rossouw, M., Victor, T., et al. (2004). SLC11A1 (NRAMP1) but not SLC11A2 (NRAMP2) polymorphisms are associated with susceptibility to tuberculosis in a high-incidence community in South Africa. *Int. J. Tuberc. Lung Dis.* 8, 1464–1471.

Ho, L., Chan, S. Y., Burk. R. D., Das, B. C., Fujinaga, K., Icenogle, J. P., et al. (1993). The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. *J. Virol.* 67, 6413–6423.

Holcombe, C. (1992). *Helicobacter pylori*: the African enigma. *Gut* 33, 429–431. doi: 10.1136/gut.33.4.429

Huibregtse, J. M., Scheffner, M., and Howley, P. M. (1993a). Cloning and expression of the cDNA for E6-AP, a protein that mediates the interaction of the human papillomavirus E6 oncoprotein with p53. *Mol. Cell. Biol.* 13, 775–784.

Huibregtse, J. M., Scheffner. M., and Howley, P. M. (1993b). Localization of the E6-AP regions that direct human papillomavirus E6 binding, association with p53, and ubiquitination of associated proteins. *Mol. Cell. Biol.* 13, 4918–4927.

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Hung, I. F., and Wong, B. C. (2009). Assessing the risks and benefits of treating *Helicobacter pylori* infection. *Therap. Adv. Gastroenterol.* 2, 141–147. doi: 10.1177/1756283X08100279

Intemann, C. D., Thye, T., Niemann, S., Browne, E. N., Amanua Chinbuah, M., Enimil, A., et al. (2009). Autophagy gene variant IRGM-261T contributes to protection from tuberculosis caused by *Mycobacterium tuberculosis* but not by *M. africanum* strains. *PLoS Pathog.* 5:e1000577. doi: 10.1371/journal.ppat.1000577

Israel, D. A., Salama, N., Krishna, U., Rieger, U. M., Atherton, J. C., Falkow, S., et al. (2001). *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14625–14630. doi: 10.1073/pnas.251551698

Jacob, F. (1977). Evolution and tinkering. *Science* 196, 1161–1166. doi: 10.1126/science.860134

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., et al. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* 41, 657–665. doi: 10.1038/ng.388

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Karlsson, E. K., Kwiatkowski, D. P., and Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* 15, 379–393. doi: 10.1038/nrg3734

Klingelhutz, A. J., and Roman, A. (2012). Cellular transformation by human papillomaviruses: lessons learned by comparing high- and low-risk viruses. *Virology* 424, 77–98. doi: 10.1016/j.virol.2011.12.018

Knolle, H. (1989). Host density and the evolution of parasite virulence. *J. Theor. Biol.* 136, 199–207. doi: 10.1016/S0022-5193(89)80226-7

Ko, D. C., and Urban, T. J. (2013). Understanding human variation in infectious disease susceptibility through clinical and cellular GWAS. *PLoS Pathog.* 9:e1003424. doi: 10.1371/journal.ppat.1003424

Kodaman, N., Pazos, A., Schneider, B. G., Piazuelo, M. B., Mera, R., Sobota, R. S., et al. (2014). Human and *Helicobacter pylori* coevolution shapes the risk of gastric disease. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1455–1460. doi: 10.1073/pnas.1318093111

Kraaijeveld, A. R., Van Alphen, J. J., and Godfray, H. C. (1998). The coevolution of host resistance and parasitoid virulence. *Parasitology* 116(Suppl. 1), S29–S45. doi: 10.1017/S0031182000084924

Lambrechts, L., Fellous, S., and Koella, J. C. (2006). Coevolutionary interactions between host and parasite genotypes. *Trends Parasitol.* 22, 12–16. doi: 10.1016/j.pt.2005.11.008

Lenski, R. E., and Levin, B. R. (1985). Constraints on the coevolution of bacteria and virulent phage – a model, some experiments, and predictions for natural communities. *Am. Nat.* 125, 585–602. doi: 10.1086/284364

Lewin, R. (1993). *Human Evolution*, 3rd Edn. Boston: Blackwell Scientific Publications.

Lewis, S. J., Baker, I., and Davey Smith, G. (2005). Meta-analysis of vitamin D receptor polymorphisms and pulmonary tuberculosis risk. *Int. J. Tuberc. Lung Dis.* 9, 1174–1177.

Libby, P., Ridker, P. M., and Maseri, A. (2002). Inflammation and atherosclerosis. *Circulation* 105, 1135–1143. doi: 10.1161/hc0902.104353

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., et al. (2000). Environmental and heritable factors in the causation of cancer – analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* 343, 78–85. doi: 10.1056/NEJM200007133430201

Little, T. J. (2002). The evolutionary significance of parasitism: do parasite-driven genetic dynamics occur ex silico? *J. Evol. Biol.* 15, 1–9. doi: 10.1046/j.1420-9101.2002.00366.x

Little, T. J., Watt, K., and Ebert, D. (2006). Parasite-host specificity: experimental studies on the basis of parasite adaptation. *Evolution* 60, 31–38. doi: 10.1111/j.0014-3820.2006.tb01079.x

Liu, L., Chen, X., Skogerbo, G., Zhang, P., Chen, R., He, S., et al. (2012). The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics* 100, 265–270. doi: 10.1016/j.ygeno.2012.07.012

Lively, C. M., and Dybdahl, M. F. (2000). Parasite adaptation to locally common host genotypes. *Nature* 405, 679–681. doi: 10.1038/35015069

Magnusson, P. K., Lichtenstein, P., and Gyllensten, U. B. (2000). Heritability of cervical tumours. *Int. J. Cancer* 88, 698–701. doi: 10.1002/1097-0215(20001201)88:5<698::AID-IJC3>3.0.CO;2-J

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140

Malik, A. N., and Godfrey-Faussett, P. (2005). Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease. *Lancet Infect. Dis.* 5, 174–183. doi: 10.1016/S1473-3099(05)01310-1

Mayerle, J., den Hoed, C. M., Schurmann, C., Stolk, L., Homuth, G., Peters, M. J., et al. (2013a). Identification of genetic loci associated with *Helicobacter pylori* serologic status. *JAMA* 309, 1912–1920. doi: 10.1001/jama.2013.4350

Mayerle, J., Kuipers, E. J., and Lerch, M. M. (2013b). Genetic variants associated with susceptibility to *Helicobacter pylori* – reply. *JAMA* 310, 976–977. doi: 10.1001/jama.2013.194772

McFall-Ngai, M., Hadfield, M. G., Bosch, T. C., Carey, H. V., Domazet-Loso, T., Douglas, A. E., et al. (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3229–3236. doi: 10.1073/pnas.1218525110

Messenger, S. L., Molineux, I. J., and Bull, J. J. (1999). Virulence evolution in a virus obeys a trade-off. *Proc. Biol. Sci.* 266, 397–404. doi: 10.1098/rspb.1999.0651

Moller, M., and Hoal, E. G. (2010). Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis (Edinb.)* 90, 71–83. doi: 10.1016/j.tube.2010.02.002

Monot, M., Honore, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., et al. (2009). Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat. Genet.* 41, 1282–1289. doi: 10.1038/ng.477

Moodley, Y., and Linz, B. (2009). *Helicobacter pylori* sequences reflect past human migrations. *Genome Dyn.* 6, 62–74. doi: 10.1159/000235763

Moodley, Y., Linz, B., Bond, R. P., Nieuwoudt, M., Soodyall, H., Schlebusch, C. M., et al. (2012). Age of the association between *Helicobacter pylori* and man. *PLoS Pathog.* 8:e1002693. doi: 10.1371/journal.ppat.1002693

Moodley, Y., Linz, B., Yamaoka, Y., Windsor, H. M., Breurec, S., Wu, J. Y., et al. (2009). The peopling of the Pacific from a bacterial perspective. *Science* 323, 527–530. doi: 10.1126/science.1166083

Munger, K., Baldwin, A., Edwards, K. M., Hayakawa, H., Nguyen, C. L., Owens, M., et al. (2004). Mechanisms of human papillomavirus-induced oncogenesis. *J. Virol.* 78, 11451–11460. doi: 10.1128/JVI.78.21.11451-11460.2004

Nicol, M. P., and Wilkinson, R. J. (2008). The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. *Trans. R. Soc. Trop. Med. Hyg.* 102, 955–965. doi: 10.1016/j.trstmh.2008.03.025

Olesen, R., Wejse, C., Velez, D. R., Bisseye, C., Sodemann, M., Aaby, P., et al. (2007). DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans. *Genes Immun.* 8, 456–467. doi: 10.1038/sj.gene.6364410

Ong, C. K., Chan, S. Y., Campo. M. S., Fujinaga, K., Mavromara-Nazos, P., Labropoulou, V., et al. (1993). Evolution of human papillomavirus type 18: an

ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. *J. Virol.* 67, 6424–6431.

Pallen, M. J., and Wren, B. W. (2007). Bacterial pathogenomics. *Nature* 449, 835–842. doi: 10.1038/nature06248

Parwati, I., van Crevel, R., and van Soolingen, D. (2010). Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect. Dis.* 10, 103–111. doi: 10.1016/S1473-3099(09)70330-5

Peek, R. M. Jr., and Blaser, M. J. (2002). *Helicobacter pylori* and gastrointestinal tract adenocarcinomas. *Nat. Rev. Cancer* 2, 28–37. doi: 10.1038/nrc703

Picard, C., Casanova, J. L., and Abel, L (2006). Mendelian traits that confer predisposition or resistance to specific infections in humans. *Curr. Opin. Immunol.* 18, 383–390. doi: 10.1016/j.coi.2006.05.005

Plummer, M., Schiffman, M., Castle, P. E., Maucort-Boulch, D., Wheeler, C. M., and ALTS Group. (2007). A 2-year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. *J. Infect. Dis.* 195, 1582–1589. doi: 10.1086/516784

Porta, C., Riboldi, E., and Sica, A. (2011). Mechanisms linking pathogens-associated inflammation and cancer. *Cancer Lett.* 305, 250–262. doi: 10.1016/j.canlet.2010.10.012

Portevin, D., Gagneux, S., Comas, I., and Young, D. (2011). Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog.* 7:e1001307. doi: 10.1371/journal.ppat.1001307

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4680–4687. doi: 10.1073/pnas.1002611107

Rector, A., Lemey, P., Tachezy, R., Mostmans, S., Ghim. S. J., Van Doorslaer, K., et al. (2007). Ancient papillomavirus–host co-speciation in Felidae. *Genome Biol.* 8, R57. doi: 10.1186/gb-2007-8-4-r57

Reed, M. B., Domenech, P., Manca, C., Su, H., Barczak, A. K., Kreiswirth, B. N., et al. (2004). A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* 431, 84–87. doi: 10.1038/nature02837

Reiling, N., Homolka, S., Walter, K., Brandenburg, J., Niwinski, L., Ernst, M., et al. (2013). Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. *MBio* 4, pii: e00250-13. doi: 10.1128/mBio.00250-13

Ridenhour, B. J., and Nuismer, S. L. (2007). Polygenic traits and parasite local adaptation. *Evolution* 61, 368–376. doi: 10.1111/j.1558-5646.2007.00029.x

Rothenbacher, D., Blaser, M. J., Bode, G., and Brenner, H. (2000). Inverse relationship between gastric colonization of *Helicobacter pylori* and diarrheal illnesses in children: results of a population-based cross-sectional study. *J. Infect. Dis.* 182, 1446–1449. doi: 10.1086/315887

Rothenbacher, D., Winkler, M., Gonser, T., Adler, G., and Brenner, H. (2002). Role of infected parents in transmission of *Helicobacter pylori* to their children. *Pediatr. Infect. Dis. J.* 21, 674–679. doi: 10.1097/00006454-200207000-00014

Rowell, J. L., Dowling, N. F., Yu, W., Yesupriya, A., Zhang, L., and Gwinn, M. (2012). Trends in population-based studies of human genetics in infectious diseases. *PLoS ONE* 7:e25431. doi: 10.1371/journal.pone.0025431

Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., and Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. U.S.A.* 97, 14668–14673. doi: 10.1073/pnas.97.26.14668

Salama, N. R., Hartung, M. L., and Muller, A. (2013). Life in the human stomach: persistence strategies of the bacterial pathogen *Helicobacter pylori*. *Nat. Rev. Microbiol.* 11, 385–399. doi: 10.1038/nrmicro3016

Samson, M., Libert, F., Doranz, B. J., Rucker, J., Liesnard, C., Farber, C. M., et al. (1996). Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 382, 722–725. doi: 10.1038/382722a0

Schiffman, M., Herrero, R., Desalle, R., Hildesheim, A., Wacholder, S., Rodriguez, A. C., et al. (2005). The carcinogenicity of human papillomavirus types reflects viral evolution. *Virology* 337, 76–84. doi: 10.1016/j.virol.2005.04.002

Schneider, B. G., Camargo, M. C., Ryckman, K. K., Sicinschi, L. A., Piazuelo, M. B., Zabaleta, J., et al. (2008). Cytokine polymorphisms and gastric cancer risk. *Cancer Biol. Ther.* 7, 157–162. doi: 10.4161/cbt.7.2.5270

Shah, S. D., Doorbar, J., and Goldstein, R. A. (2010). Analysis of host–parasite incongruence in papillomavirus evolution using importance sampling. *Mol. Biol. Evol.* 27, 1301–1314. doi: 10.1093/molbev/msq015

Sharma, S. K., and Mohan, A. (2004). Extrapulmonary tuberculosis. *Ind. J. Med. Res.* 120, 316–353.

Shi, Y., Hu, Z., Wu, C., Dai, J., Li, H., Dong, J., et al. (2011). A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat. Genet.* 43, 1215–1218. doi: 10.1038/ng.978

Shi, Y., Li, L., Hu, Z., Li, S., Wang, S., Liu, J., et al. (2013). A genome-wide association study identifies two new cervical cancer susceptibility loci at 4q12 and 17q12. *Nat. Genet.* 45, 918–922. doi: 10.1038/ng.2687

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., and Alm., E, J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244. doi: 10.1038/nature10571

Sokurenko, E. V., Gomulkiewicz, R., and Dykhuizen, D. E. (2006). Source-sink dynamics of virulence evolution. *Nat. Rev. Microbiol.* 4, 548–555. doi: 10.1038/nrmicro1446

Storey, A., Thomas, M., Kalita, A., Harwood, C., Gardiol, D., Mantovani, F., et al. (1998). Role of a p53 polymorphism in the development of human papillomavirus-associated cancer. *Nature* 393, 229–234. doi: 10.1038/30400

Stucki, D., and Gagneux, S. (2013). Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. *Tuberculosis (Edinb.)* 93, 30–39. doi: 10.1016/j.tube.2012.11.002

Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., et al. (1998). Free recombination within *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* 95, 12619–12624. doi: 10.1073/pnas.95.21.12619

Susser, M., and Stein, Z. (2002). Civilization and peptic ulcer. *Int. J. Epidemiol.* 31, 13–17. doi: 10.1093/ije/31.1.13

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437, 1299–1320. doi: 10.1038/nature04226

Thompson, J. N. (2014). Natural selection, coevolution, and the web of life. *Am. Nat.* 183, iv–v. doi: 10.1086/674238

Thompson, J. N., Nuismer, S. L., and Gomulkiewicz, R. (2002). Coevolution and maladaptation. *Integr. Comp. Biol.* 42, 381–387. doi: 10.1093/icb/42.2.381

Thye, T., Niemann, S., Walter, K., Homolka, S., Intemann, C. D., Chinbuah, A., et al. (2011). Variant G57E of mannose binding lectin associated with protection against tuberculosis caused by *Mycobacterium africanum* but not by *M. tuberculosis*. *PLoS ONE* 6:e20908. doi: 10.1371/journal.pone.0020908

Thye, T., Owusu-Dabo, E., Vannberg, F. O., van Crevel, R., Curtis, J., Sahiratmadja, E., et al. (2012). Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet.* 44, 257–259. doi: 10.1038/ng.1080

Thye, T., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., Gyapong, J., et al. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* 42, 739–741. doi: 10.1038/ng.639

Tiemersma, E. W., van der Werf, M. J., Borgdorff, M. W., Williams, B. G., and Nagelkerke, N. J. (2011). Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. *PLoS ONE* 6:e17601. doi: 10.1371/journal.pone.0017601

Tornesello, M. L., Duraturo, M. L., Salatiello, I., Buonaguro, L., Losito. S., Botti, G., et al. (2004). Analysis of human papillomavirus type-16 variants in Italian women with cervical intraepithelial neoplasia and cervical cancer. *J. Med. Virol.* 74, 117–126. doi: 10.1002/jmv.20154

Torres, J., Correa, P., Ferreccio, C., Hernandez-Suarez, G., Herrero, R., Cavazza-Porro, M., et al. (2013). Gastric cancer incidence and mortality is associated with altitude in the mountainous regions of Pacific Latin America. *Cancer Causes Control* 24, 249–256. doi: 10.1007/s10552-012-0114-8

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel. B. L., Duncan. A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Vaezi, M. F., Falk, G. W., Peek, R. M., Vicari, J. J., Goldblum, J. R., Perez-Perez, G. I., et al. (2000). CagA-positive strains of *Helicobacter pylori* may protect against Barrett's esophagus. *Am. J. Gastroenterol.* 95, 2206–2211. doi: 10.1111/j.1572-0241.2000.02305.x

Van Doorslaer, K. (2013). Evolution of the papillomaviridae. *Virology* 445, 11–20. doi: 10.1016/j.virol.2013.05.012

Velez, D. R., Hulme, W. F., Myers, J. L., Stryjewski, M. E., Abbate, E., Estevan, R., et al. (2009). Association of SLC11A1 with tuberculosis and interactions with

NOS2A and TLR2 in African-Americans and Caucasians. *Int. J. Tuberc. Lung Dis.* 13, 1068–1076.

Wheeler, C. M. (2008). Natural history of human papillomavirus infections, cytologic and histologic abnormalities, and cancer. *Obstet. Gynecol. Clin. North Am.* 35, 519–536; vii. doi: 10.1016/j.ogc.2008.09.006

White, A. E., Livanos, E. M., and Tlsty, T. D. (1994). Differential disruption of genomic integrity and cell cycle regulation in normal human fibroblasts by the HPV oncoproteins. *Genes Dev.* 8, 666–677. doi: 10.1101/gad.8.6.666

Wilfert, L., and Schmid-Hempel, P. (2008). The genetic architecture of susceptibility to parasites. *BMC Evol. Biol.* 8:187. doi: 10.1186/1471-2148-8-187

Wirth, T., Wang, X., Linz, B., Novick, R. P., Lum, J. K., Blaser, M., et al. (2004). Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4746–4751. doi: 10.1073/pnas.0306629101

Woolhouse, M. E., Webster, J. P., Domingo, E., Charlesworth, B., and Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* 32, 569–577. doi: 10.1038/ng1202-569

Wroblewski, L. E., Peek, R. M. Jr., and Wilson, K. T. (2010). *Helicobacter pylori* and gastric cancer: factors that modulate disease risk. *Clin. Microbiol. Rev.* 23, 713–739. doi: 10.1128/CMR.00011-10

Xi, L. F., Koutsky, L. A., Galloway, D. A., Kuypers, J., Hughes, J. P., Wheeler, M., et al. (1997). Genomic variation of human papillomavirus type 16 and risk for high grade cervical intraepithelial neoplasia. *J. Natl. Cancer Inst.* 89, 796–802. doi: 10.1093/jnci/89.11.796

zur Hausen, H. (1989). Papillomavirus in anogenital cancer: the dilemma of epidemiologic approaches. *J. Natl. Cancer Inst.* 81, 1680–1682. doi: 10.1093/jnci/81.22.1680

zur Hausen, H. (1991). Viruses in human cancers. *Science* 254, 1167–1173. doi: 10.1126/science.1659743

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 June 2014; paper pending published: 30 June 2014; accepted: 05 August 2014; published online: 25 August 2014.

*Citation: Kodaman N, Sobota RS, Mera R, Schneider BG and Williams SM (2014) Disrupted human–pathogen co-evolution: a model for disease. Front. Genet. **5**:290. doi: 10.3389/fgene.2014.00290*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 02 April 2014
doi: 10.3389/fgene.2014.00068

A cautionary note on ignoring polygenic background when mapping quantitative trait loci via recombinant congenic strains

J Concepción Loredo-Osti*

Department of Mathematics and Statistics, Memorial University, St. John's, NL, Canada

Edited by:
José M. Âlvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Hongying Dai, Children's Mercy Hospital, USA
Carl Nettelblad, Uppsala University, Sweden

****Correspondence:***
J Concepción Loredo-Osti, Department of Mathematics and Statistics, Memorial University, Henrietta Harvey Building, St. John's, NL A1C 5S7, Canada
e-mail: jcloredoosti@mun.ca

In gene mapping, it is common to test for association between the phenotype and the genotype at a large number of loci, i.e., the same response variable is used repeatedly to test a large number of non-independent and non-nested hypotheses. In many of these genetic problems, the underlying model is a mixed model consistent of one or very few major genes concurrently with a genetic background effect, usually thought as of polygenic nature and, consequently, modeled through a random effects term with a well-defined covariance structure dependent upon the kinship between individuals. Either because the interest lies only on the major genes or to simplify the analysis, it is habitual to drop the random effects term and use a simple linear regression model, sometimes complemented with testing via resampling as an attempt to minimize the consequences of this practice. Here, it is shown that dropping the random effects term has not only extreme negative effects on the control of the type I error rate, but it is also unlikely to be fixed by resampling because, whenever the mixed model is correct, this practice does not allow to meet some basic requirements of resampling in a gene mapping context. Furthermore, simulations show that the type I error rates when the random term is ignored can be unacceptably high. As an alternative, this paper introduces a new bootstrap procedure to handle the specific case of mapping by using recombinant congenic strains under a linear mixed model. A simulation study showed that the type I error rates of the proposed procedure are very close to the nominal ones, although they tend to be slightly inflated for larger values of the random effects variance. Overall, this paper illustrates the extent of the adverse consequences of ignoring random effects term due to polygenic factors while testing for genetic linkage and warns us of potential modeling issues whenever simple linear regression for a major gene yields multiple significant linkage peaks.

Keywords: misspecified genetic models, bootstrapping mixed models, recombinant congenic strains, ignoring random effects, mapping quantitative trait loci

1. INTRODUCTION

For more than four decades, linear mixed models have been used in a wide range of applications because of their conceptual simplicity and flexibility to accommodate correlated sources of variation as well as fixed regressors. A generic linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices, $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients, $\boldsymbol{\gamma}$ is a vector of random effects, and $\mathbf{e}$ is the vector of errors. It is also common to assume that $\boldsymbol{\gamma}$ and $\mathbf{e}$ are independent and both have null expectation and finite variances. In many situations, either intentionally or unintentionally, the statistical analysis is carried out ignoring the term $\mathbf{Z}\boldsymbol{\gamma}$ in the model. This practice, although recognized as inefficient, has been thought to be harmless whenever the interest resides solely on a subset of the regression coefficients with the remaining parameters of the model deemed as nuisance. This thought seems to be mostly based on the fact that $\boldsymbol{\beta}^o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is still an unbiased and consistent estimator of $\boldsymbol{\beta}$. However, it is well known that ignoring $\mathbf{Z}\boldsymbol{\gamma}$ and using ordinary least squares, results in an estimator of $\mathrm{Var}(\boldsymbol{\beta}^o)$ that is biased and inconsistent as well as non-independent of $\boldsymbol{\beta}^o$ (Dhymes, 1978). Of course, this will affect the distribution properties associated with $\boldsymbol{\beta}^o$ under normality or, otherwise, the asymptotic properties of its distribution. It has been suggested that this problem can be mitigated if testing is done through resampling. However, the adverse consequences of dropping the random term from the mixed model is unlikely to be fixed by the use of resampling methods. In this paper, a specific application to genetic mapping via recombinant congenic strains (RCS) of experimental animals is used to illustrate this. Briefly speaking, genetic mapping can be seen as a problem in which the association of one dependent variable (the phenotype) with a large number of potential explicative variables (the marker genotypes) is tested one-by-one or by taking a very small number of markers at once. An RCS panel is a replicable mapping population for which animals within the same strain are considered to be genetically identical and related to different degrees

with animals from other strains. Such an inter-strain relationship results in what is known as the genetic background effect and, whenever this effect is understood as the result of the addition of many components of minuscule effect, the inclusion of a random effects term in the model would be the natural way to account for it.

A mouse panel of RCS is obtained by mating mice from two genetically distinct inbred strains (a donor strain and a recipient strain) followed by two or more rounds of backcrossing to the recipient strain and subsequent sister × brother mating without selection for particular markers or phenotypes for a minimum of 20 generations. The genetic resolution of the panel is controlled by the number of backcrossing rounds. Because of this construction, each strain of an RCS panel can be thought of as an inbred strain in which segments of random length from the genome of a recipient strain have been replaced with the corresponding segments from a donor strain. The main consequence of this breeding scheme is that non-linked genes controlling the same trait are separated and fixed in haplotypes of different strains, allowing the possibility of studying them individually. The standard RCS panel uses two backcross generations and, consequently, the total length of the segments from recipient strain constitute on average the 87.5% of the genome of each strain; the remaining 12.5% represents the total expected length of the replaced genome segments. Without loss of generality, this is the type of RCS considered in this paper. For a more comprehensive description of the RCS and their use in gene mapping see Démant and Hart (1986), Moen et al. (1992), and Fortin et al. (2001b, 2007). Once the RCS panel have been established, the whole panel is genotyped to obtain full characterization of the genome of each strain. Each genotype data set can then be used for the analysis of all individuals of the same strain; this is an important money-saving feature of the design since it does not require of re-genotyping each individual because, except for *de novo* mutations, all pups from the same strain are genetically identical.

Although most mouse geneticists agree that RCS are a powerful resource to map loci associated with complex traits, there is some disagreement on how to do the analysis. Originally, when the use of RCS for genetic mapping was proposed, the core idea was to look into the stain distribution pattern with respect to a phenotype of interest and identify the strain that exhibited the largest deviation from the other strains in the RCS panel and subsequently cross it with the recipient strain to obtain F_1 and F_2 progenies to be analyzed by standard methods (Démant and Hart, 1986; Fortin et al., 2001b). Two examples of the application of this approach are reported in Fortin et al. (2001a) and Müllerová and Hozák (2004). The problem is that contrasting phenotypes from F_1 mice versus the ones from the recipient strain will only be effective for dominant traits, while the power for additive traits will be diminished and lost completely for recessive traits. On the other hand, the analysis of the F_2 mice requires new genotyping, which not only defeats the economic advantages of having developed RCS, but more importantly, because every F_2 individual has different genotype, this approach is not suited for complex quantitative traits when a single measurement may not be reliable enough to determine the phenotype (Moen et al., 1992). Alternatively, there is a designs consisting of taking a sample of mice from each strain and analyzing the whole panel together. Although this approach does not require additional genotyping and has the potential for making more efficient use of the phenotypic variation, also opens more room for analysis pitfalls if the proper model is not used. For example, Joober et al. (2002) uses a QTL mapping procedure equivalent to simple linear regression at the markers ignoring genetic background which, as pointed by Palmer and Airey (2003), it may result in false positive rates far in excess of the nominal value, even when Bonferroni corrections are used. Another common way to address the problem is to use strain averages as the phenotype and treat the panel of means as a backcross dataset for analysis purposes. This is essentially the "interval mapping" procedure proposed by Shao et al. (2010) and equivalent to the one used by Thifault et al. (2008). This approach may substantially reduce the power for RCS panels with reduced number of strains and it does not deal with the fact that the strains, related because their background, may not have the same kinship degree at genomic level and consequently the phenotype means may be not only non-independent but heteroscedastic, as well. Lee et al. (2006) and Camateros et al. (2010) extend the simple linear regression to account for the genetic background by adding a fixed factor ("background proportion" in the first paper; "background indicator" in the second). Although better than ignoring the background, from the genetics standpoint, it is difficult to justify the plausibility of a fixed effects model under the assumption that the background effect is the result of the additive action of many genes of minuscule effect. In fact, I argue that the natural way to model such a background effect consistent with the principles outlined by Fisher (1919) is through the inclusion of a random effects term in the model as implemented in Di Pietrantonio et al. (2010). In this paper, I describe in detail a procedure for the analysis of a quantitative trait locus (QTL) that models the genetic background (assumed to be of polygenic nature) as a random effect term and use this to show how the omission of such a term in the model leads to conclusions that are wrong and inconsistent with the data.

2. MODELS

2.1. THE NAIVE QTL MODEL FOR AN RCS PANEL

In its simplest form, at each marker position m, $m = 1, 2, \ldots, M$, the RCS/QTL model for the ith individual, $i = 1, 2, \ldots, n$, can be written as

$$y_i = \mu + q_{im}\,\xi_m + e_i \tag{2}$$

where y_i denotes the phenotype for the ith individual, ξ_m denotes the major locus effect associated with the mth marker, q_{im} is the indicator of the BB genotype at the mth position which is determined by the RCS data, and the e_is are a set of independent random variables with distribution $\mathcal{N}(0, \sigma^2)$ (AA and BB are the genotypes of the donor and recipient parental strain, respectively). Of course, under an oligogenic model, at most, a handful of ξ_ms should be different from zero. In fact, it is common practice that at the first screening, the estimation is carried out by regression at each marker under the assumption of only one major gene. When the presumption of a dense enough genotyping marker panel is not correct, procedures like modified interval

mapping can be used instead. Variations of the problem include conditioning on a given set of markers. The salient feature of this design is that, at the mth marker position, one looks across the RCS panel and classifies each strain as either AA or BB, since under the model (Equation 2), this is the only source of genetic variation when estimating ξ_m. However, this model ignores the fact that individuals from the same strain are genetically identical (assuming no new mutation at the locus under scrutiny), and strains with the same ancestral background share large portions of their genome so that even without the involvement of a major gene, there is more likely to be reduced variation within strains. In a nutshell, regression mapping works by testing the association of the phenotype with the observed genotype at each marker location so that finding significant linkage at any position implies testing the M null hypotheses, $\xi_m = 0$. Clearly, most of these hypotheses as well as their test statistics are not independent. This may lead to problems in the control of the type I error rate if multiple testing is not addressed properly. Another irregularity results from the fact that with a dense genotyping panel the number of tested hypotheses can by far exceed the sample size. Because of these considerations, p-value estimation by resampling of residuals has been seen as a plausible alternative. For this paper, the problem is addressed through bootstrap.

2.1.1. Computation of p-values

The estimation of genome-wide corrected p-values by resampling requires that under the null hypothesis: (i) each resample is taken from an exchangeable distribution, (ii) the variation of the original sample is preserved through all resamples, and (iii) the genome-wide baseline for the test statistics at each position is the same. The first two requirements are standard for resampling in regression (Davison and Hinkley, 1997; Anderson and Ter Braak, 2003). The last requirement is imposed to ensure that the uncorrected p-values across the genome are comparable (this is particularly important when there are missing genotype data). One way to estimate corrected p-values is to select an ensemble of test statistics whose marginal distribution is the same when the model does not contain any major locus.

Since under model (Equation 2) and the hypothesis of no major gene, the distribution of $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is exchangeable, resampling from the raw observations will also preserve the variation through the pseudo-observations. This means that in the absence of non-genetic regressors or other non-oligogenic factors, resampling the raw phenotypes either by permutation or through bootstrap will produce similar results. Furthermore, under these premises, basic sampling and hypothesis testing principles indicate that a permutation based procedure will be more efficient and powerful. However, this is not necessarily the case when the premises are removed. Should the model also contain fixed non-genetic regressors, resampling from the leverage-adjusted residuals under the null hypothesis would be a procedure that approximates exchangeability while preserving the original variation of the data. However, under this situation, resampling from leverage-adjusted residuals results in a procedure with acceptable properties only in the bootstrap case (Davison and Hinkley, 1997), while this is not longer guaranteed when resampling via permutation. The main issue is that sampling without replacement magnifies the effects of modest departures from exchangeability. Then, permuting leverage-adjusted residuals may not be good enough (even worst, it may not be valid) and we would require of a much more elaborate and computer intensive procedure to obtain residuals guaranteed to be at least weakly exchangeable so that permutation works properly (see, for example, Kherad-Pajouh and Renaud, 2010). To complete the requirements listed above regarding the possibility of missing genotypes, we propose to use the test statistic defined by the expression

$$z_m = t_m \left(1 - \frac{1}{4\nu_m}\right)\left(1 + \frac{t_m^2}{2\nu_m}\right)^{-\frac{1}{2}} \quad \text{where} \quad t_m = \frac{|\hat{\xi}_m|}{\hat{\sigma}_{\hat{\xi}_m}} \tag{3}$$

and $\hat{\xi}_m$ is the ordinary least squares estimate of ξ_m, $m = 1, 2, \ldots, M$, i.e., z_m is just t_m, our familiar t-statistic with ν_m degrees of freedom, transformed into a z-score (ν_m may vary slightly from marker to marker due to missing data). Another option would be a modified t-statistic t'_m in which the mth estimate of variance s_m^2 used to compute $\hat{\sigma}^2_{\hat{\xi}_m}$ is replaced by s_0^2, the estimate under the null hypothesis. With no missing genotypes the use of any of z_m, t'_m, and t_m would yield approximately the same p-value estimates.

2.1.2. Bootstrap procedure for simple linear regression at the markers

The following bootstrap procedure computes the genome-wide corrected p-values for model (Equation 2) with the test statistic (Equation 3):

STEP 1. At each marker position, m, fit the simple linear regression at the markers model (Equation 2), use (Equation 3) to compute the test statistic z_m, and obtain the genome-wide set of statistics $\mathcal{Z}_M = \{z_m,\ m = 1, 2, \ldots, M\}$. Also, set the genome-wide acceptance count vector to zero.

STEP 2. Sample with replacement from the raw vector of phenotypes, $\mathbf{y} \in \mathbb{R}^n$, to obtain $\mathbf{y}^* \in \mathbb{R}^n$, a bootstrapped full replica of $\mathbf{y}$, and use this vector to compute $z^*_{\max} = \max\{z^*_m\}$, where z^*_m, $m = 1, 2, \ldots, M$, is the test statistic at the mth locus, computed by using $\mathbf{y}^*$, the vector of the pseudo-observations, instead of the original vector of phenotypes.

STEP 3. For each z_m in $\mathcal{Z}_M$, if $z_m \leq z^*_{\max}$, add a unit to the mth entry of the acceptance count vector.

STEP 4. Repeat steps 1 and 2 R times and then compute the estimate of the vector of p-values by dividing the acceptance count vector by R.

This resampling scheme can be seen as an adaptation of a regular regression residuals bootstrapping procedure (Davison and Hinkley, 1997), coupled with Roy's union-intersection principle (Roy, 1953) to control for the genome-wide type I error rate. When applied to the analysis of the RCS panel, this procedure is valid when there is only one observation per strain or when the within-strain variation is negligible. Otherwise, a random term in the model has been neglected and, regardless of $\hat{\xi}_m$ being an

unbiased estimator of ξ_m, the exchangeability requirement cannot be met and the most likely consequence would be an inflated type I error rate. In fact, as per arguments given by Churchill and Doerge (1994) and Churchill and Doerge (2008), this statement is correct not only for the bootstrap and RCS, but also for permutation test procedures applied to any study design involving replicable mapping populations because, as for bootstrap, the Fisher (1935) principle of permutation also relies on exchangeability. For simple experimental designs such as an intercross or a backcross mating, the individual units can safely be assumed to be exchangeable. However, it would be wrong to assume exchangeability for more complicated designs, like advanced intercross, heterogeneous stocks and RCS.

2.2. THE QTL MIXED MODEL FOR AN RCS PANEL

The previous simple linear model (Equation 2) generalizes to a model of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{q}_m\xi_m + \mathbf{e} \tag{4}$$

where $\mathbf{y}$ represents the phenotype vector, $\mathbf{q}_m$ is a vector with each entry being an indicator variable of the genotype BB at the marker position m with ξ_m being its associated effect (major gene effect), $\boldsymbol{\gamma}$ is a random effects vector associated with the genetic background with $\mathrm{E}(\boldsymbol{\gamma}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\gamma}) = \sigma^2_\gamma\,\boldsymbol{\Delta}_1$, with $\sigma^2_\gamma > 0$ and $\boldsymbol{\Delta}_1$, a positive-definite matrix, both assumed to be constant, although unknown, $\mathbf{X}$ is a matrix of fixed covariates and its corresponding parameter vector $\boldsymbol{\beta}$, $\mathbf{e}$ is a vector of independent and identically distributed random variables representing the error term with $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2\,\mathbf{I}$. Up to a multiplicative constant, $\boldsymbol{\Delta}_1$ is a function of the length of the segments identical by descent shared amongst strains. For an established RCS panel there are only two possible identity states between pairs of strains at a given locus: either (i) all four alleles are identical by descent ($\boldsymbol{\Delta}_1$ is the matrix holding the pairwise probabilities for this state), or (ii) the strains have different allelic forms and thus identical by descent only amongst themselves. So an estimator of $\boldsymbol{\Delta}_1$ with "a high degree of precision" can be reached. Such an estimator uses only genomic information and does not involve $\mathbf{y}$, so when estimating the parameters, one can assume that $\boldsymbol{\Delta}_1$ is given. Another option is to take the entries of $\boldsymbol{\Delta}_1$ as the expected value of the proportion of the genome shared identical by descent between the respective strains under the RCS panel construction described above, i.e.,

$$\delta_{1ij} = \begin{cases} 1 & \text{if } i = j \\ \frac{15}{16} & \text{if } i \text{ and } j \text{ have the same background} \\ \frac{1}{16} & \text{if } i \text{ and } j \text{ have different backgrounds.} \end{cases} \tag{5}$$

This option, although not the most efficient, does capture the main features of the design and yields a variance structure for the random effects vector that can be exploited in the implementation of the resampling algorithm. For example, if all the strains in the panel under scrutiny have the same background and the simplified expectation-based $\boldsymbol{\Delta}_1$ is used, then the distribution of the vector of random effects is exchangeable. Nonetheless, replacing a genomic-based $\boldsymbol{\Delta}_1$ estimate by its theoretical expectation (Equation 5) implies ignoring important information regarding the correlation of the additive polygenic effects associated to the genetic background.

2.2.1. Estimation

The estimation for the mixed linear model has been extensively discussed in the literature (Harville, 1977; Henderson, 1986). Here we develop an application of these standard methods to the RCS design. Without loss of generality, let us consider the linear mixed model (Equation 1) with $\mathrm{Var}(\boldsymbol{\gamma}) = \sigma^2_\gamma\,\boldsymbol{\Delta}_1$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$. Thus

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{Var}(\mathbf{y}) = \sigma^2\left(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{I}\right) = \sigma^2\,\boldsymbol{\Sigma}$$

where $\mathbf{G} = \lambda\boldsymbol{\Delta}_1$ and $\lambda = \frac{\sigma^2_\gamma}{\sigma^2}$, i.e., λ represents the signal-to-noise ratio. Under the assumption of no major gene and only polygenic background, λ is related to the heritability coefficient. When $\mathbf{G}$ is known, the best linear unbiased estimator of $\boldsymbol{\beta}$ and the best linear unbiased predictor of $\boldsymbol{\gamma}$ (also known as a shrinkage estimator) can be written as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{v} \quad \text{and} \quad \hat{\boldsymbol{\gamma}} = \mathbf{G}\mathbf{Z}'\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{v} - \mathbf{W}\tilde{\boldsymbol{\beta}}),$$

respectively, where $\mathbf{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}$ and $\mathbf{v} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{y}$. Also

$$\hat{\sigma}^2 = \frac{1}{N - \mathrm{rank}(\mathbf{W})}(\mathbf{v} - \mathbf{W}\tilde{\boldsymbol{\beta}})'(\mathbf{v} - \mathbf{W}\tilde{\boldsymbol{\beta}})$$

$$\hat{\sigma}^2_\gamma = \frac{1}{\mathrm{rank}(\mathbf{G})}\left(\hat{\boldsymbol{\gamma}}'\mathbf{G}^{-1}\hat{\boldsymbol{\gamma}} + \hat{\sigma}^2\mathrm{tr}(\mathbf{G}^{-1}\mathbf{C})\right)$$

with

$$\mathbf{C} = (\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{G}^{-1})^{-1} \quad \text{and} \quad \mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Notice that the previous expressions cannot be computed unless the signal-to-noise ratio, λ, is known. A situation of a more practical interest is an iterative procedure on which λ is replaced by its estimate and, once that the estimates of σ^2 and σ^2_γ have been updated, a refinement of the estimate of λ is obtained and so on. This iterative procedure will result in a $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ that are no longer linear, nonetheless, they preserve most of the desirable properties present in their linear counterpart (Jiang, 1998).

2.2.2. Mixed model resampling scheme

Let us now focus our attention toward a resampling scheme appropriate for RCS data under a mixed model. By now, it is obvious that the bootstrap procedure described in the previous section will not work for the mixed model (Equation 4). A crude extension to this procedure would consist of computing

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}$$

and resampling from $\hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{e}}$ to obtain $\boldsymbol{\gamma}^*$ and $\mathbf{e}^*$ so that the pseudo-observation $\mathbf{y}^*$ could be recovered as

$$\mathbf{y}^* = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{e}^*.$$

However, it is straightforward to see that these residuals are not exchangeable and they are biased toward zero. Thus, they may not adequately represent the hypothesis tested nor reflect the true variation of the model.

Alternatively, note that when $\boldsymbol{\beta}$ and λ are known, it follows from the model under the null hypothesis that $\mathrm{E}(\mathbf{v}) = \mathbf{W}\boldsymbol{\beta}$ and $\mathrm{Var}(\mathbf{v}) = \sigma^2\,\mathbf{I}$ which implies that the distribution of the vector of residuals, $\boldsymbol{\epsilon} = \mathbf{v} - \mathbf{W}\boldsymbol{\beta}$, is exchangeable. This suggests the following residuals resampling scheme:

(1) given $\tilde{\lambda}$ and $\tilde{\boldsymbol{\beta}}$ obtained under the mixed model without a major gene, i.e., under the null hypothesis, compute $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\mathbf{W}}$ by replacing λ with $\tilde{\lambda}$ and $\boldsymbol{\Delta}_1$ with its genomic-based estimate; then, obtain the leverage-adjusted residuals

$$\tilde{\boldsymbol{\epsilon}} = \mathbf{D}(\tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{y} - \tilde{\mathbf{W}}\tilde{\boldsymbol{\beta}})$$

where $\mathbf{D}$ is a diagonal matrix with each of the non-zero elements given by $(1 - h_{ii})^{-1}$ and h_{ii} is the ith leverage coefficient;

(2) with replacement, resample from $\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^n$ to obtain $\boldsymbol{\epsilon}^* \in \mathbb{R}^n$, its bootstrapped replica, and construct the vector of pseudo-observations as

$$\mathbf{v}^* = \tilde{\mathbf{W}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^*.$$

If instead of a bootstrap procedure based on leverage-adjusted residuals we want to use a residuals-based permutation procedure, then we need to extend the method of Kherad-Pajouh and Renaud (2010) to get weak exchangeability of residuals. However, when λ is estimated from the data, such an extension is not possible and we would have to rely on approximations. More research is needed to explore this direction.

Outside of a genetics context, there is a number of permutation and bootstrap procedures for mixed models whose objective is testing the components of variance (for example, Fitzmaurice et al., 2007; Sinha, 2009; Lee and Braun, 2012; Samuh et al., 2012). However, they cannot be applied in our case because we are interested in the regression coefficients (or a subset of them) and the variance of the random effects is just nuisance parameter. Incidentally, when testing the components of variance, bootstrap has the edge over most permutation procedures (Samuh et al., 2012).

2.2.3. Bootstrap procedure for the mixed linear model

According to the foregoing argument, generalization to the previous bootstrap procedure to compute the genome-wide corrected p-values for the mixed model (Equation 4) goes as follows:

STEP 0. Compute $\boldsymbol{\Delta}_1$ from the genotype data of the RCS panel, and under the null hypothesis, obtain $\tilde{\lambda}$, $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\mathbf{W}}$ and $\tilde{\boldsymbol{\epsilon}}$ as described in (i) above.

STEP 1. At each marker position, m, fit the model

$$\tilde{\mathbf{v}} = \begin{pmatrix} \tilde{\mathbf{W}} & \tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{q}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \xi_m \end{pmatrix} + \boldsymbol{\epsilon}. \qquad (6)$$

Of course, this model is equivalent to model (Equation 4), the RCS/QTL mixed model, with λ replaced by $\tilde{\lambda}$. Compute the model parameter estimates with the outlined mixed model procedure as well as the test statistic set $\mathcal{Z} = \{z_m,\ m = 1, 2, \ldots, M\}$ by using Equations (6) and (3); set the acceptance count vector to zero.

STEP 2. Draw a pseudo-observation $\mathbf{v}^*$ by using the proposed resampling scheme in (ii) above and fit the major gene model in model (Equation 6) with $\tilde{\mathbf{v}}$ replaced by $\mathbf{v}^*$ to obtain the set of bootstrapped test statistics $\{z_m^*\}$ and its associated critical value $z_{\max}^* = \max\{z_m^*\}$.

STEP 3. For each z_m in $\mathcal{Z}$, if $z_m \leq z_{\max}^*$, add a unit to the mth entry of the acceptance count vector.

STEP 4. Repeat steps 2 and 3 R times and compute the p-value estimates by dividing the acceptance count vector by R.

To my knowledge, this bootstrap procedure for the analyzing a panel of RCS has not been proposed before Di Pietrantonio et al. (2010) and this paper contains the first detailed derivation and study of its properties. In fact, the resampling methods (mostly conditional permutation) applied to analyze RCS have not used mixed models, but consider the strain effect as fixed which is inconsistent with the hypothesis of a genetic background of polygenic nature or discard information by using only the estimated strain means (for example, Gill and Boyle, 2005; Thifault et al., 2008; Camateros et al., 2010).

3. RESULTS

One straightforward way to show the effect of ignoring the random effects term in a mixed model is by simulation. The idea is to generate a dataset from a model that includes a random term for genetic background and noise, but is free of any major locus. Then compare the p-value profiles (actually, $-\log_{10} p$ profiles) obtained by the use of the naive model (Equation 2) as well as the mixed model (Equation 4). For this simulation study, the genotypes of an RCS panel of 36 strains that were described in Fortin et al. (2001b) were used. The panel originally had 37 lines and 625 microsatellite markers; since then, one line has died out and six markers were removed for reliability reasons. Although a much larger set of single nucleotide polymorphism markers for this RCS panel is also available, I think that this set of 619 markers is enough to show the harmful effects of fitting the wrong model on the inference. Of course, more markers will only exacerbate the problem. For this simulation experiment, six different values for the signal-to-noise ratio parameter λ were chosen (0, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, and 2). Under a standard additive polygenic model, i.e., a model without major genes, the signal-to-noise parameter is a function of the heritability coefficient (the chosen values correspond to the heritability proportions of 0, $\frac{1}{9}$, $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{2}{3}$, respectively). In every simulation run, a sample of seven individuals from each strain was simulated under the assumption of no major gene, i.e., under model (Equation 4) with $\xi_m = 0$ for all markers, $m = 1, 2, \ldots, M$. The value of σ^2 was fixed for all simulations to 1.175, while $\mathbf{X}\boldsymbol{\beta}$ was fixed as a vector with 7 in all its entries. Simulations for each value of λ were run 1000 times and both methodologies, the mixed model as well as the bootstrapped naive regression at the markers were applied to the simulated datasets with 10, 000 as the number of resamples for every dataset.

In gene mapping studies, a significant peak is defined as the most extreme point of a region beyond the p-value threshold according to some pre-specified genome-wide type I error rate (Churchill and Doerge, 1994). For this study, we use a value of 0.01 or equivalently, a threshold value of 2 on a $-\log_{10} p$-scale. **Tables 1–3** summarize the results of these simulations. As expected, whenever there is not a polygenic term in the model (i.e., $\lambda = 0$), both methodologies produce identical results. However, the picture changes when $\lambda > 0$. In this case, it is quite obvious that ignoring the random effects term has pernicious consequences even for modest levels of λ, the signal-to-noise ratio, while the proposed mixed model method keeps the genome-wide type I error rate relatively close to the nominal value. However, the empirical type I error rates obtained by the proposed procedure seem to increase slightly with λ (**Table 3**). This phenomenon may be due to the fact that the makers used for mapping purposes are also used to estimate the probability of identity by descent between strains and, to a lesser extent, the fact that the the bootstrap procedure is based on residuals computed with λ and $\boldsymbol{\beta}$ estimated from the same data. Nonetheless, the moral of this exercise is that whenever simple regression of a major gene model produces many significant peaks, a warning flag about the model validity should be raised.

Table 1 | Percentage of declared significant peaks with a bootstrap genome-wide adjusted significance level of 0.01 when the proposed mixed model methodology is used.

%		Signal-to-noise ratio (λ)					
		0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**
Number of significant peaks	0	99.2	98.7	98.9	98.3	98.5	98.4
	1	0.8	0.5	0.5	0.7	0.7	0.4
	2	0	0.2	0.1	0.2	0.5	0.3
	3	0	0.1	0	0.4	0.1	0.3
	4	0	0	0.3	0.1	0.1	0.1
	5	0	0.1	0.1	0.1	0	0.1
	6+	0	0.2	0.1	0.3	0.1	0.4

Estimates based on 1000 simulated datasets for each λ.

Table 2 | Percentage of declared significant peaks with a bootstrap genome-wide adjusted significance level of 0.01 when a naive regression at the markers is used.

%		Signal-to-noise ratio (λ)					
		0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**
Number of significant peaks	0	99.2	61.1	47.3	38.8	23.9	17.2
	1	0.8	3.9	5.1	5.2	8.1	7.1
	2	0	3.7	4.1	5.1	4.9	6.0
	3	0	1.5	3.5	3.1	3.0	5.3
	4	0	2.5	4.6	3.3	2.1	4.1
	5	0	2.1	5.3	3.2	2.9	2.4
	6+	0	25.2	30.1	41.3	55.1	57.9

Estimates based on 1000 simulated datasets for each λ.

The histogram of a typical dataset obtained by simulation from a model with polygenic effects only would look like the one shown in **Figure 1**. Nonetheless, for this histogram I chose a dataset for which simple linear regression produces a very large number of significant peaks. If a major locus were at play, one would expect to have a well-defined bimodal distribution, so this histogram seems consistent with the generating model of no major gene. However, when we look into the p-value profiles obtained through the model that ignores the genetic background term, instead of profiles consistent with the model we will have something extreme as shown by dashed lines in **Figure 2**. According to the profiles on this figure, one might conclude that all chromosomes have at least one significant peak, fact that does not appear to be supported by the histogram of the data, and more conclusively, this is in conflict with the generating model. If anything, it can be argued that the data distribution may seem a bit skewed, but one may expect that estimation of p-values via bootstrapping of residuals should not be too sensitive to this. Of course, as for bi-modality, skewness may also be caused by a mixture of distributions. However, a very strong peak, as any of the ones spotted on every chromosome, is difficult to conceive without a conspicuous bimodal distribution. Even with the use of robust regression estimates instead of the obtained by regular least squares to minimize the potential impact of outliers on the estimation, these profiles change very little (data not shown). When the missing random effects term is introduced into the model (solid blue

Table 3 | Empirical genome-wide type I error rates obtained via bootstrap in the simulation study (0.01 is the nominal value and the number of simulated datasets for each λ is 1000).

	Signal-to-noise ratio (λ)					
	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	**1**	**2**
Naive regression	0.008	0.389	0.527	0.612	0.761	0.808
Mixed model	0.008	0.013	0.011	0.017	0.015	0.016

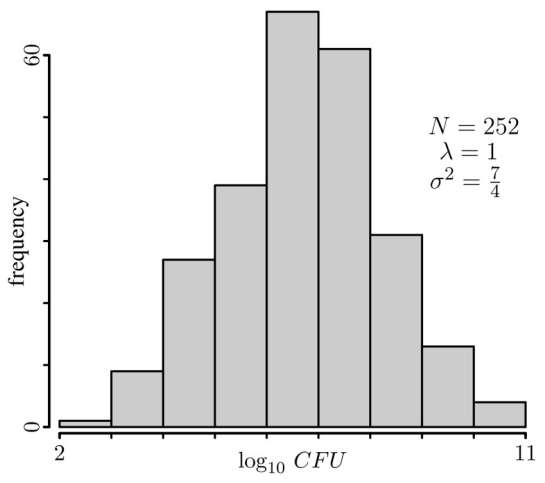


FIGURE 1 | Typical histogram of simulated data. The p-value profiles of the data on this histogram were computed and plotted in **Figure 2**.

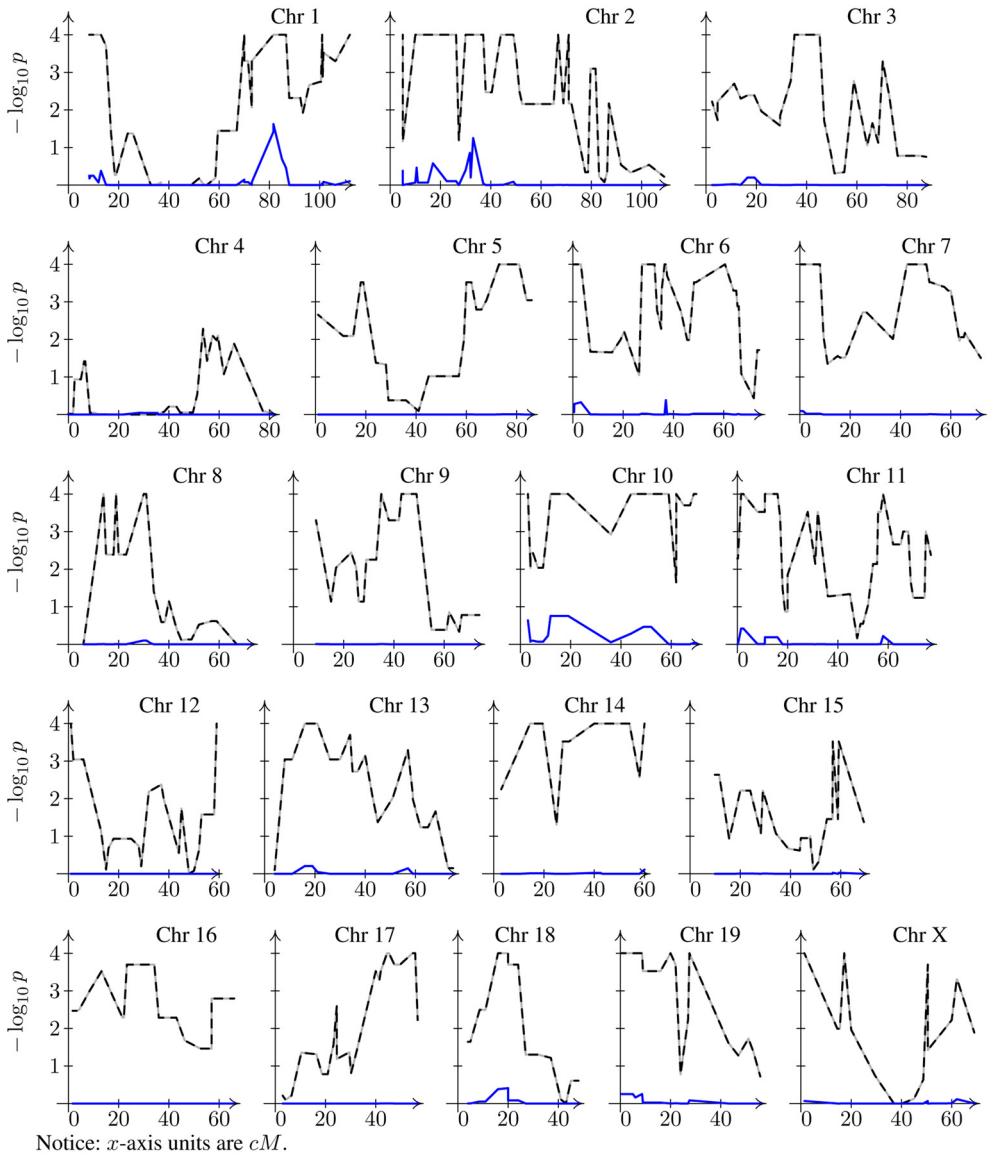


FIGURE 2 | Bootstrap genome-wide corrected *p*-value profiles. Dashed line for naive model (Equation 2) and solid line (sometimes hardly distinguishable from the *x*-axis line) for the mixed model (Equation 6). Note that both profiles have been corrected for multiple testing.

lines in **Figure 2**), *p*-value profiles become consistent with the generating model. Repetition of this exercise on any other simulated datasets yields similar results, although the specific resulting profiles most likely are not be the same.

4. DISCUSSION

This paper proposes a bootstrapping procedure to estimate the *p*-values under a mixed model applied to gene mapping when RCS are used. The method can be easily adapted for other replicable mapping population/designs. This procedure is a generalization of the linear regression bootstrap of residuals coupled with the union-intersection principle aimed to control the genome-wide type I error rate. A simulation study with different values of the signal-to-noise ratio unequivocally shows that when a panel of RCS is used for mapping, ignoring one random effects term in a mixed linear model can have pernicious consequences, resulting in inflated type I error rates and leading to the declaration of significant linkage peaks were no such peaks should be found. The simulation study also shows that the proposed bootstrap procedure seems to produce slightly inflated type I error rates as the signal-to-noise ratio increases. This problem is likely due to the fact that the markers used for mapping are also used to estimate the length of the segments shared identical by descent but also it can be associated with a stronger departure from exchangeability as the ratio increases. In any case, the problem deserves further scrutiny. The proposed bootstrap procedure for mixed models is quite general and can easily be adapted to non-genetic problems.

FUNDING

This work has been supported by the Canadian Institutes of Health Research.

ACKNOWLEDGMENTS

The author expresses gratitude to M. Fujiwara, E. Schurr, T. di Pietrantonio, and K. Morgan for the discussion and comments that substantially improved the manuscript.

REFERENCES

Anderson, M. J., and Ter Braak, C. J. F. (2003). Permutation tests for multi-factorial analysis of variance. *J. Statist. Comput. Simul.* 73, 85–113. doi: 10.1080/00949650215733

Camateros, P., Marino, R., Fortin, A., Martin, J. G., Skamene, E., Sladek, R., et al. (2010). Identification of novel chromosomal regions associated with airway hyperresponsiveness in recombinant congenic strains of mice. *Mamm. Genome* 21, 28–38. doi: 10.1007/s00335-009-9236-z

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Churchill, G. A., and Doerge, R. W. (2008). Naive application of permutation testing leads to inflated type I error rates. *Genetics* 178, 609–610. doi: 10.1534/genetics.107.074609

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511802843

Démant, P., and Hart, A. A. M. (1986). Recombinant congenic strains–a new tool for analyzing genetic traits determined by more than a gene. *Immunogenetics* 24, 416–422. doi: 10.1007/BF00377961

Dhymes, P. J. (1978). *Introductory Econometrics*. New York, NY: Springer. doi: 10.1007/978-1-4612-6292-3

Di Pietrantonio, T., Hernandez, C., Girard, M., Verville, A., Orlova, M., Belley, A., et al. (2010). Strain-specific differences in the genetic control of two closely related mycobacteria. *PLoS Pathog.* 6:e1001169. doi: 10.1371/journal.ppat.1001169

Fisher, R. A. (1919). The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinb.* 52, 399–433. doi: 10.1017/S0080456800012163

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* 63, 942–946. doi: 10.1111/j.1541-0420.2007.00775.x

Fortin, A., Cardon, L. R., Tam, M., Skamene, E., Stevenson, M. M., and Gros, P. (2001a). Identification of a new malaria susceptibility locus (Char4) in recombinant congenic strains of mice. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10793–10798. doi: 10.1073/pnas.191288998

Fortin, A., Diez, E., Henderson, J. E., Mogil, J. S., Gros, P., and Skamene, E. (2007). "Decoding the genomic control of immune reactions: novartis foundation symposium 281," in *Decoding the Genomic Control of Immune Reactions: Novartis Foundation Symposium 281*, eds G. Bock and J. Goode (Chichester: John Wiley), 141–155. doi: 10.1002/9780470062128

Fortin, A., Diez, E., Rochefort, D., Laroche, L., Malo, D., Rouleau, G. A., et al. (2001b). Recombinant congenic strains derived from A/J and C57BL/6J: a tool for genetic dissection of complex traits. *Genomics* 74, 21–35. doi: 10.1006/geno.2001.6528

Gill, K. J., and Boyle, A. E. (2005). Quantitative trait loci for novelty/stress-induced locomotor activation in recombinant inbred (ri) and recombinant congenic (rc) strains of mice. *Behav. Brain Res.* 161, 113–124. doi: 10.1016/j.bbr.2005.01.013

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998

Henderson, C. R. (1986). Recent developments in variance and covariance estimations. *J. Anim. Sci.* 63, 208–216.

Jiang, J. (1998). Asymtotic properties of the empirical BLUP and BLUE in mixed linear models. *Stat. Sinica* 8, 861–885.

Joober, R., Zarate, J.-M., Rouleau, G.-A., Skamene, E., and Boksa, P. (2002). Provisional mapping of quantitative trait loci modulating the acoustic startle response and prepulse inhibition of acoustic startle. *Neuropsychopharmacology* 27, 765–781. doi: 10.1016/S0893-133X(02)00333-0

Kherad-Pajouh, S., and Renaud, O. (2010). An exact permutation method for testing any effect in balancedand unbalanced fixed effect ANOVA. *Comput. Stat. Data Anal.* 5, 1881–1893. doi: 10.1016/j.csda.2010.02.015

Lee, O. E., and Braun, T. M. (2012). Permutation tests for random effects in linear mixed models. *Biometrics* 68, 486–493. doi: 10.1111/j.1541-0420.2011.01675.x

Lee, P. D., Ge, B., Greenwood, C. M., Sinnett, D., Fortin, Y., Brunet, S., et al. (2006). Mapping cis-acting regulatory variation in recombinant congenic strains. *Physiol. Genomics* 25, 294–302. doi: 10.1152/physiolgenomics.00168.2005

Moen, C. J., Groot, P. C., Dietrich, W., Stoye, J. P., Lander, E. S., and Démant, P. (1992). The recombinat congenic strains for analysis of multigenic traits: genetic composition. *FASEB J.* 6, 2806–2835.

Müllerová, J., and Hozák, P. (2004). Use of recombinant congenic strains in mapping disease-modifying genes. *News Physiol. Sci.* 19, 105–109. doi: 10.1152/nips.01512.2003

Palmer, A. A., and Airey, D. C. (2003). Inappropriate choice of the experimental unit leads to a dramatic overestimation of the significance of quantitative trait loci for prepulse inhibition and startle response in recombinant congenic mice. *Neuropsychopharmacology* 28, 818. doi: 10.1038/sj.npp.1300064

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat.* 24, 220–238. doi: 10.1214/aoms/1177729029

Samuh, M. H., Grilli, L., Rampichini, C., Salmaso, L., and Lunardon, N. (2012). The use of permutation tests for variance components in linear mixed models. *Commun. Stat. Theor. Methods* 41, 3020–3029. doi: 10.1080/03610926.2011.587933

Shao, H., Sinasac, D. S., Burrage, L. C., Hodges, C. A., Supelak, P. J., Palmert, M. R., et al. (2010). Analyzing complex traits with congenic strains. *Mamm. Genome* 21, 276–286. doi: 10.1007/s00335-010-9267-5

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Can. J. Stat.* 37, 219–234. doi: 10.1002/cjs.10012

Thifault, S., Sun, Y., Fortin, A., Skamene, E., Lalonde, R., Tremblay, J., et al. (2008). Genetic determinants of emotionality and stress response in AcB/BcA recombinant congenic mice and *in silico* evidence of convergence with cardiovascular candidate genes. *Hum. Mol. Genet.* 17, 331–344. doi: 10.1093/hmg/ddm277

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2013; accepted: 17 March 2014; published online: 02 April 2014.
*Citation: Loredo-Osti JC (2014) A cautionary note on ignoring polygenic background when mapping quantitative trait loci via recombinant congenic strains. Front. Genet. **5**:68. doi: 10.3389/fgene.2014.00068*
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 20 February 2014
doi: 10.3389/fgene.2014.00032

A modified generalized Fisher method for combining probabilities from dependent tests

Hongying Dai[1,2,3]*, J. Steven Leeder[2,4] and Yuehua Cui[5]

[1] Department of Pediatrics, Research Development and Clinical Investigation, Children's Mercy Hospital, Kansas City, MO, USA
[2] Department of Pediatrics, University of Missouri-Kansas City, Kansas City, MO, USA
[3] Department of Informatic Medicine and Personalized Health, University of Missouri-Kansas City, Kansas City, MO, USA
[4] Department of Pediatrics, Clinical Pharmacology and Therapeutic Innovation, Children's Mercy Hospital, Kansas City, MO, USA
[5] Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

Edited by:
José M. Álvarez-Castro, Universidade de Santiago de Compostela, Spain

Reviewed by:
Wei Hou, Stony Brook University, USA
J. Concepcion Loredo-Osti, Memorial University, Canada

***Correspondence:**
Hongying Dai, Department of Pediatrics, Research Development and Clinical Investigation, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO 64108, USA
e-mail: hdai@cmh.edu

Rapid developments in molecular technology have yielded a large amount of high throughput genetic data to understand the mechanism for complex traits. The increase of genetic variants requires hundreds and thousands of statistical tests to be performed simultaneously in analysis, which poses a challenge to control the overall Type I error rate. Combining *p*-values from multiple hypothesis testing has shown promise for aggregating effects in high-dimensional genetic data analysis. Several *p*-value combining methods have been developed and applied to genetic data; see Dai et al. (2012b) for a comprehensive review. However, there is a lack of investigations conducted for dependent genetic data, especially for weighted *p*-value combining methods. Single nucleotide polymorphisms (SNPs) are often correlated due to linkage disequilibrium (LD). Other genetic data, including variants from next generation sequencing, gene expression levels measured by microarray, protein and DNA methylation data, etc. also contain complex correlation structures. Ignoring correlation structures among genetic variants may lead to severe inflation of Type I error rates for omnibus testing of *p*-values. In this work, we propose modifications to the Lancaster procedure by taking the correlation structure among *p*-values into account. The weight function in the Lancaster procedure allows meaningful biological information to be incorporated into the statistical analysis, which can increase the power of the statistical testing and/or remove the bias in the process. Extensive empirical assessments demonstrate that the modified Lancaster procedure largely reduces the Type I error rates due to correlation among *p*-values, and retains considerable power to detect signals among *p*-values. We applied our method to reassess published renal transplant data, and identified a novel association between B cell pathways and allograft tolerance.

Keywords: generalized Fisher method (Lancaster procedure), weight function, correlated *p*-values, multiple hypothesis testing, high dimensional genetic data

INTRODUCTION

Rapid developments in molecular technology have created high throughput data in search of genetic variants associated with complex traits. As the cost of experiments goes down, the amount of data that can be generated, and the resulting complexity of statistical analysis required to interpret the data goes up. The increase of genetic variants requires more statistical testing to be performed simultaneously, which poses a challenge to control the genome wide Type I error rate. False discovery rate (FDR) and its extended methods have been proposed to adjust *p*-values in multiple tests in order to control the genome wide Type I error (Benjamini and Hochberg, 1995; Cheng and Pounds, 2007). However, in large-scale hypothesis testing, these methods often require very a large sample size to maintain power of detecting risk factors.

The global test (also named omnibus test) of *p*-values can combine evidence and turn dimensionality from a curse into rich information. From a systems biology perspective, genes, cells, tissues, and organs function as a system through metabolic networks and cell signaling networks. In non-Mendelian inheritance patterns, such as complex disorders, a subset of genetic variants may jointly confer moderate effects in mediating molecular activities. As a result, signals may not be significant in single marker-single trait analysis, but many such values from related genes might provide valuable information on gene function and regulation. For instance, in pathway analysis (Khatri et al., 2012) and gene set enrichment analysis (Subramanian et al., 2005), multiple genes that work together to serve a particular biological function are often analyzed jointly as a gene set. Several pathway repositories, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004), PANTHER classification system for protein sequence data (Nikolsky and Bryant, 2009), and Reactome pathways in humans (Matthews et al., 2009) have been established, and are continually being updated. For non-Mendelian diseases and complex traits, identification of isolated genetic variants is insufficient to summarize the complex

association with disease. The "most-significant SNPs/genes" approach often detects variants with small effect sizes and odds ratios ranging between 1.3 and 2 (Wacholder et al., 2004). Therefore, integrating information from pathways, gene sets, and networks will provide useful information in understanding the gene regulation mechanism. Furthermore, filtration techniques can be integrated with global testing of *p*-values to remove sets of genetic variants that are not related to traits, and thereby reduce the dimensionality of the data (Dai and Charnigo, 2008; Dai et al., 2012a).

The global test of *p*-values evaluates the pattern (distribution) of *p*-values instead of selecting *p*-values less than an arbitrary threshold. Therefore, this method has the potential to identify multiple genes with small effects. If we assume that all individual tests are independent and arise from genetic variants with no effects, then *p*-values are identically and independently distributed as Uniform (0, 1). Taking this as a null hypothesis for the pattern of *p*-values in the global test, one can assess whether *p*-values, especially small *p*-values, are generated by chance. The global test of *p*-values is robust and can be applied to *p*-values from varying statistical models including *t*-tests, analysis of variance (ANOVA), linear mixed models, and so forth. Multiple simulation studies and case studies have demonstrated that this approach usually has sufficient power to detect signals of genetic association from a group of genes. For instance, Peng et al. (2010) has assessed Fisher's combination test and Sidak's combination test, Sime's combination test and the FDR method using 13 published genome wide association studies (GWAS), and the results indicate that combined *p*-value approaches can identify biologically meaningful pathways associated with the disease susceptibility. A review of methods of global test of *p*-values, developmental trends and their application to genetic data analysis has been presented by (Dai et al., 2012b).

One category of global tests of *p*-values involves combining *p*-values in the form of $\sum_i H(p_i)$, where *p*-values might first be transformed by a function H. So far, several statistical methods have been developed to combine *p*-values. Let $p_i (i = 1, 2, \ldots, n)$ be independent *p*-values obtained from n hypothesis tests. Under the null hypothesis (H_0) that *p*-values follow a Uniform (0, 1) distribution, Fisher (1932) shows that $-2\sum_{i=1}^{n} \ln(p_i)$ follows a chi-square distribution with $2n$ degrees of freedom. For a one sided test with a nominal error rate of α, one can reject the null hypothesis when the test statistics exceeds the $(1-\alpha)$*100% percentile of χ^2_{2n}. Stouffer (Stouffer et al., 1949) proposed a *z*-test by transforming *p*-values to standard normal variables, i.e., $\sum_{i=1}^{n} \frac{\Phi^{-1}(1-p_i)}{\sqrt{n}}$, where Φ^{-1} is the inverse Cumulative Distribution Function (CDF) for $N(0, 1)$. Under the null hypothesis, the *z*-test statistic follows $N(0, 1)$.

Although there is no consensus regarding the most powerful method of combining *p*-values, Littell and Folks (1971, 1973) demonstrated that the Fisher's method of combining independent tests is asymptotically Bahadur efficient (Bahadur, 1967). Subsequently, weighting schemes have been incorporated into the Fisher's method and the *z*-test. Lancaster (1961) generalized the Fisher method by converting independent *p*-values to chi-square variables with w_i degrees of freedom and he showed that $\sum_{i=1}^{m} \gamma^{-1}_{(w_i/2,2)}(1-p_i) \sim \chi^2_d$, $d = \sum_i w_i$ under H_0, where $\gamma^{-1}_{(w_i/2,2)}$ is the inverse CDF of Gamma distribution. Mosteller and Bush (1954) proposed a weighted *z*-test, $\sum_i w_i \Phi^{-1}(1-p_i)/\sqrt{\sum_i w_i^2}$, which follows $N(0, 1)$ under H_0.

In a separate paper, we have proved that the Lancaster procedure achieves the optimal Bahadur efficiency. We further demonstrated that the Lancaster procedure yields higher Bahadur efficiency than the weighted *z*-test. The Bahadur efficiency ratio gives the limiting ratio of sample sizes required by two statistics to attain an equally small significance level. Thus, Bahadur efficiency is an important method to compare test statistics. From the perspective of Bahadur efficiency, the Lancaster procedure asymptotically requires a relatively smaller sample size than other weighted *p*-value combining methods. This prompted us to focus on modification of the Lancaster procedure for correlated genetic data in this work.

Although the Fisher's method and Lancaster procedure both achieve the optimal Bahadur efficiency, the Lancaster procedure is more general and can be viewed as a generalized Fisher's method with weighting functions. There are three advantages to carefully select appropriate weight functions in genetic data analysis. Firstly, weight functions allow incorporation of prior biological information. Genetic data are complex and can be measured from different sources. Thus, weight functions can be used as a tool to incorporate meaningful information from different sources in order to interpret and derive biological insight from gene expression profiles. (Wu and Lin, 2009) provides a review of statistical methods for analysis of microarray data by incorporating prior biological knowledge using gene sets and biological pathways, which consist of groups of biologically similar genes. They show that the use of prior knowledge has led to a better understanding of the biological mechanisms underlying phenotypic responses. Secondly, weight functions can be used to remove bias. For instance, larger genes may contain more probes and/or SNPs. Therefore, larger genes will exert a stronger influence on the *p*-value combining methods as compared to smaller genes (Wang et al., 2007). To avoid this bias, one can consider a weight function to adjust for gene size when combining *p*-values. We will illustrate this approach in sections Empirical Assessments and Case Study: Renal Transplant Tolerance Data. Thirdly, as suggested by Benjamini and Hochberg (1997), Genovese et al. (2006), procedures that assign weights positively associated with the underlying alternative hypotheses will usually improve power. Therefore, one needs to carefully choose an appropriate weight function, either based on the biological knowledge, or by statistical hypotheses. An arbitrary weight is inappropriate for the Lancaster procedure.

In this work, we will provide modifications to the Lancaster procedure to accommodate correlation structures among *p*-values. The proposed method provides a generalization to the Fisher's method with a weight function and can be used in pathway analysis and gene sets enrichment analysis for a variety of genetic data including microarray gene expression data, GWAS data, and next generation sequencing data. In essence, investigators first dissect genetic variants by biological functions

or prior knowledge, then combine the p-values from these gene sets to identify whether a proportion of genetic variants are associated with traits.

CORRELATED LANCASTER PROCEDURES

In this section, we allow p-values to be correlated. Consider a Lancaster test statistic $T = \sum_{i=1}^{n} \gamma^{-1}_{(w_i/2,2)}(1-p_i)$ where $\gamma^{-1}_{(w_i/2,2)}$ is the inverse CDF of Gamma distribution with a shape parameter $w_i/2$ and a scale parameter 2. This transformation converts $p_i \sim \text{Uniform}(0,1)$ to a chi-square distribution, i.e., $\gamma^{-1}_{(w_i/2,2)}(1-p_i) \sim \chi^2_{w_i}$ where $\chi^2_{w_i}$ is a chi-square distribution with $w_i > 0$ degree(s) of freedom. The parameter w_i serves as a weight function to adjust the individual p-values. When p-values are independent, T has an exact chi-square distribution with $\sum_{i=1}^{n} w_i$ degrees of freedom.

For correlated p-values, $T = \sum_{i=1}^{n} \gamma^{-1}_{(w_i/2,2)}(1-p_i)$ does not follow $\chi^2_{\sum_{i=1}^{n} w_i}$. The distribution of T does not have an explicit analytical form. To address this issue, we consider a Satterthwaite approximation by approximating a scaled T statistic with a new chi-square distribution (Li et al., 2011). Let $cT \approx \chi^2_\nu$ where $c > 0$ is a scalar and $\nu > 0$ is the degree of freedom for the approximated chi-square distribution. Note that

$$
\begin{aligned}
E(T) &= E\left(\sum_{i=1}^{n} \gamma^{-1}_{(w_i/2,2)}\left(1-p_i\right)\right) = \sum_{i=1}^{n} w_i \text{ and} \\
\text{Var}(T) &= \text{var}\left(\sum_{i=1}^{n} \gamma^{-1}_{(w_i/2,2)}\left(1-p_i\right)\right) \\
&= \sum_{i=1}^{n} \text{var}\left(\gamma^{-1}_{(w_i/2,2)}\left(1-p_i\right)\right) \\
&\quad + 2\sum_{i<j} \text{cov}\left(\gamma^{-1}_{(w_i/2,2)}\left(1-p_i\right), \gamma^{-1}_{(w_i/2,2)}\left(1-p_j\right)\right) \\
&= 2\sum_{i=1}^{n} w_i + 2\sum_{i<j} \rho_{ij},
\end{aligned}
$$

where $\rho_{ij} = \text{cov}\left(\gamma^{-1}_{(w_i/2,2)}(1-p_i), \gamma^{-1}_{(w_i/2,2)}\left(1-p_j\right)\right)$ takes the correlations among p-values into account.

We propose the following five approaches to approximate the distribution of T. In approximation (A), we use the Satterthwaite method to match the mean and variance of cT and χ^2_ν, and then solve the equations to derive c and ν. Koziol (1996) have proposed multiple methods to approximate the Lancaster procedure, but these approximations require the assumption of independence. In approximation (B)–(E), we extend the work of Koziol (1996) to correlated data by first approximating cT with χ^2_ν then approximating χ^2_ν using varying methods.

- T_A approximation.
  Correlation among p-values is taken into consideration, and then Satterthwaite's approximation is used (Patnaik, 1949) to derive new degrees of freedom:

$$
T_A = cT \approx \chi^2_\nu, \text{ where } c = \frac{\nu}{E(T)} \text{ and } \nu = 2\frac{[E(T)]^2}{\text{var}(T)}.
$$

- T_B approximation.
  cT is first approximated by χ^2_ν, followed by Fisher's approximation (Fisher, 1922) to χ^2_ν:

$$
T_B = \sqrt{2\frac{\nu T}{E(T)}} \approx N(\sqrt{2\nu-1}, 1).
$$

- T_c approximation.
  After approximating cT by χ^2_ν, the Wilson–Hilferty approximation is performed (Wilson and Hilferty, 1931) to derive χ^2_ν.

$$
\text{Let } T_c = \sqrt[3]{\frac{T}{E(T)}}, \text{ then } T_c \approx N\left(1-2/(9\nu), \sqrt{2/(9\nu)}\right).
$$

- T_D approximation.
  Approximate cT by χ^2_ν, followed by the Cornish–Fisher expansion (Fisher and Cornish, 1960) to χ^2_ν. Let x_α denote the α-percentage point of the standard normal distribution, that is, $\Phi(x_\alpha) = \alpha$. It follows that the corresponding percentage point for $T_D = \frac{\nu T}{E(T)}$ is given by

$$
\begin{aligned}
&\nu + \sqrt{2\nu}x_\alpha + \frac{2}{3}(x_\alpha^2 - 1) + \frac{x_\alpha^3 - 7x_\alpha}{9\sqrt{2\nu}} - \frac{6x_\alpha^4 + 14x_\alpha^2 - 32}{405\nu} \\
&\quad + \frac{9x_\alpha^5 + 256x_\alpha^2 - 433x_\alpha}{4860\nu\sqrt{2\nu}}.
\end{aligned}
$$

- T_E approximation.
  Approximate cT by χ^2_ν then perform saddle point approximation (Lugannani and Rice, 1980) to χ^2_ν. Let $T_E = \frac{T}{E(T)}$. Then $\Pr(Y_E \le y) = \Phi(a_y) - \phi(b_y^{-1} - a_y^{-1})$ for $y \neq 1$ and $\Pr(Y_E \le 1) = 0.5 - (3\sqrt{\pi\nu})^{-1}$, where $a_y = \sqrt{2\nu(yt_y - K(t_y))}\text{sign}(t_y)$, $b_y = t_y\sqrt{\nu K''(t_x)}$ and $K(t) = -0.5\log(1-2t)$, and $t_y = (y-1)/2y$.

When the covariance ρ_{ij} is unknown, one can use the permutation approach to estimate ρ_{ij} by shuffling the phenotype variable among subjects. For the kth permutation ($k = 1, 2, \ldots, m$), we keep the genetic variants within the subject to preserve the correlation structure, then randomly assign the phenotype variable to subjects. Individual hypothesis testing can be done on all n genetic variants separately to generate the p-value vector $p^k = (p_1^k, p_2^k, \ldots p_n^k)^t$. The permutation is repeated $m = 1000$ times, and ρ_{ij} is estimated from $(p^1, p^2, \ldots p^m)$.

The accuracy of the five approximate distributions to the correlated Lancaster procedure is then assessed using p-values with varying correlation structures. We consider six different types of correlation structures, including fixed and random compound symmetric as well as random positive definite variance-covariance structures for Σ. Let I be an identity matrix, $\vec{1}$ be a vector of 1 s, $\otimes$ be the Kronecker product, and superscript t

be the transposition. In Cases I–V, let $\Sigma = \text{Block} \otimes I_{20}$ be compound symmetric variance matrices with 20 blocks of size 5 where $\text{Block} = \vec{1}_5\vec{1}_5^t\rho + (1-\rho)I_5$. We vary ρ over two fixed values with $\rho = 0.3$ for moderate dependence and $\rho = 0.6$ for strong dependence. In addition, we simulate random correlation coefficients from beta and uniform distributions, i.e., $\rho \sim \beta(0.3, 1.5)$ and $\rho \sim \text{uniform}(-0.2, 0.2)$, which ensures that 20 variance blocks have distinct correlation coefficients ρ within Σ. More generally, we consider random positive definite correlation matrices Σ that vary across samples and simulation runs.

The quantile-quantile (Q-Q) plot assessing the accuracy of the proposed methods when the correlation coefficient $\rho = 0.3$ is shown in **Figure 1**. For clarity, the Lancaster statistic T that combines n p-values is renamed as $T_n^{\text{Lancaster}}$ in **Figure 1**. For the original Lancaster procedure under the independence assumption, the general trend of the Q-Q plot is flatter than the reference line $y = x$, indicating the limiting distribution for the test statistic in the original Lancaster procedure is less dispersed than the distribution of $T_n^{\text{Lancaster}}$ under correlation structures. As a result, the original Lancaster procedure will have severely inflated Type I errors. In contrast, the five approximations $(T_A, \ldots, T_E)$ match the underlying distribution of $T_n^{\text{Lancaster}}$. For data with stronger internal correlation, T_A, T_D, and T_E better approximate $T_n^{\text{Lancaster}}$. The Q-Q plots under other correlation structures are similar to **Figure 1**. To save space, these similar results are not shown, but can be provided upon request.

EMPIRICAL ASSESSMENTS

We assess the Type I error rates and power for the proposed correlated Lancaster procedures and compare them to the independent Lancaster procedure (Lancaster, 1961). SNPs from a pathway of haploid GWAS are simulated using linkage disequilibrium (LD) (Li et al., 2011). Let q_1 and q_2 be the minor allele frequencies (MAFs) at loci 1 and 2. Assuming Hardy–Weinberg equilibrium, the genotype at locus 1 can be randomly generated using a binomial distribution. Given the distribution of SNP at locus 1, one can simulate the genotype at locus 2. To do so, let D be a measure of LD. Then the conditional probability for the genotype at locus 2 given the genotype at locus 1 can be expressed as $P(A|B) = [q_A q_B + D]/q_B$, $P(a|B) = [(1-q_A)q_B - D]/q_B$, $P(A|b) = [q_A(1-q_B) - D]/(1-q_B)$, and $P(a|b) = [(1-q_A)(1-q_B) + D]/(1-q_B)$ where A and B represent the minor alleles at the two loci. For a diploid genome, similar idea can be applied and the simulation details can be found at Cui et al. (2008). We simulate a pathway with 5 genes with varying numbers of SNPs in each gene listed in parenthesis i.e., G1(12), G2(8), G3(5), G4(3), G5(2). The MAF of each SNP was set to be 0.3. We simulate different levels of LD for SNPs from

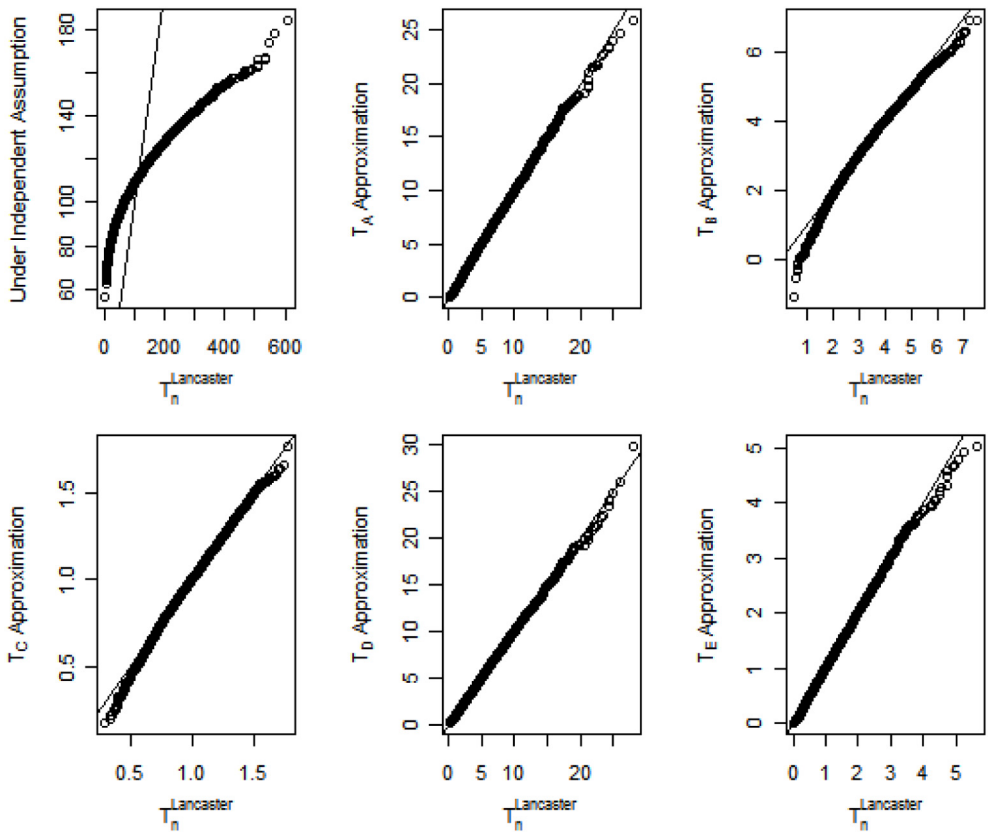


FIGURE 1 | Q-Q plots for distributions of the Lancaster statistic when *p*-values are correlated with correlation coefficient $\rho = 0.3$.

the same gene with $D = 0$, 1.5, 2, and uniform(0, maximum of LD). The variable $D = 0$, 1.5, and 2 suggests no LD, moderate LD, and very strong LD among SNPs with the corresponding correlation $R = 0$, 0.71, and 0.95. Six scenarios for disease susceptibility (p) are simulated

- Case I: $\ln(p/(1-p)) = \beta_1 G_{1,2} + \beta_2 G_{1,5} + \beta_3 G_{1,7} + \beta_4 G_{1,8} + \beta_5 G_{1,12}$.
- Case II: $\ln(p/(1-p)) = \beta_1 G_{2,2} + \beta_2 G_{2,4} + \beta_3 G_{2,6} + \beta_4 G_{3,2} + \beta_5 G_{3,3}$.
- Case III: $\ln(p/(1-p)) = \beta_1 G_{3,2} + \beta_2 G_{3,4} + \beta_3 G_{4,1} + \beta_4 G_{4,3} + \beta_5 G_{5,1}$.
- Case IV: $\ln(p/(1-p)) = \beta_1 G_{1,1} + \beta_2 G_{1,3} + \beta_3 G_{1,7} + \beta_4 G_{1,8} G_{1,10} G_{1,11} + \beta_5 G_{1,12}$.
- Case V: $\ln(p/(1-p)) = \beta_1 G_{3,1} + \beta_2 G_{3,3} + \beta_3 G_{4,2} + \beta_4 G_{3,2} G_{3,4} + \beta_5 G_{4,3} G_{5,1}$.
- Case VI: $\ln(p/(1-p)) = \beta_1 G_{1,2} + \beta_2 G_{2,2} + \beta_3 G_{3,3} + \beta_4 G_{5,2} + \beta_5 G_{1,5} G_{1,7} + \beta_6 G_{3,3} G_{5,1}$.

Weight functions can be used to remove potential bias when combining p-values. Wang et al. (2007) and others have noted that larger genes contain more probes and/or SNPs. Therefore, larger genes may exert a stronger influence on the p-value combining methods compared to smaller genes. To avoid this bias, we set the weight function $w_i = 2/\sqrt{n_i}$ where n_i is the number of SNPs in the ith gene. When $n_i = 1$, $\gamma^{-1}_{(w_i/2,\,2)}(1 - p_i)$ transforms p-value into a variable with χ^2_2 distribution.

We simulate data with sample sizes $n = 200$ (**Tables 1**, **4**) and $n = 400$ (**Tables 2**, **3**), respectively. For simplicity, we assume the same effect size for all of the regression coefficients. For each set of data, we perform the original and modified Lancaster procedures to assess the pathway data by combining p-values from individual tests. We set nominal error rate to be 0.05. The simulation is repeated 1000 times.

Due to LD, SNPs from the same gene are correlated. We first assess the Type I error rate of the test statistics by testing H_0 :

Table 1 | Type I error and power for independent Lancaster Procedure and five approximations to correlated Lancaster Procedures when sample size = 200 and linkage disequilibrium $D = 0.15$.

	Independent Lancaster procedure	T_A	T_B	T_C	T_D	T_E
CASE I						
$\beta = 0$	*0.101*	0.038	0.042	0.039	0.039	0.038
$\beta = 0.4$	0.999	0.995	0.995	0.995	0.995	0.995
$\beta = 0.6$	1	1.000	1	1	1	1
CASE II						
$\beta = 0$	*0.1*	0.037	0.041	0.038	0.038	0.037
$\beta = 0.4$	0.947	0.863	0.875	0.864	0.865	0.863
$\beta = 0.6$	0.997	0.995	0.995	0.995	0.995	0.995
CASE III						
$\beta = 0$	*0.078*	0.038	0.038	0.038	0.038	0.038
$\beta = 0.4$	0.735	0.506	0.522	0.508	0.507	0.506
$\beta = 0.6$	0.961	0.864	0.876	0.866	0.866	0.863
CASE IV						
$\beta = 0$	*0.107*	0.046	0.051	0.046	0.047	0.046
$\beta = 0.4$	0.997	0.997	0.997	0.997	0.997	0.997
$\beta = 0.6$	1	1	1	1	1	1
CASE V						
$\beta = 0$	*0.084*	0.036	0.038	0.037	0.037	0.036
$\beta = 0.4$	0.884	0.71	0.724	0.71	0.711	0.71
$\beta = 0.6$	0.989	0.952	0.957	0.953	0.953	0.952
CASE VI						
$\beta = 0$	*0.084*	0.036	0.038	0.037	0.037	0.036
$\beta = 0.4$	0.741	0.57	0.585	0.572	0.572	0.568
$\beta = 0.6$	0.953	0.898	0.904	0.898	0.898	0.898

A weight function is applied to adjust for the gene size.*

**The nominal error rate is set to be 0.05. Type I error rates are listed when $\beta = 0$. Power is listed when $\beta > 0$. Inflated Type I error rates are italicized.*

***A weight function $w_i = 2/\sqrt{n_i}$ is applied to each test to adjust for the size of gene.*

Table 2 | Type I error and power for independent Lancaster Procedure and five approximations to correlated Lancaster Procedures when sample size = 400 and linkage disequilibrium $D = 0.20$.

	Independent Lancaster procedure	T_A	T_B	T_C	T_D	T_E
CASE I						
$\beta = 0$	*0.13*	0.051	0.052	0.051	0.051	0.051
$\beta = 0.4$	1	1	1	1	1	1
$\beta = 0.6$	1	1	1	1	1	1
CASE II						
$\beta = 0$	*0.134*	0.05	0.051	0.05	0.05	0.05
$\beta = 0.4$	0.999	0.997	0.998	0.998	0.998	0.997
$\beta = 0.6$	1	1	1	1	1	1
CASE III						
$\beta = 0$	*0.116*	0.045	0.048	0.045	0.045	0.045
$\beta = 0.4$	0.986	0.908	0.915	0.908	0.908	0.908
$\beta = 0.6$	1	0.998	0.998	0.998	0.998	0.998
CASE IV						
$\beta = 0$	*0.109*	0.046	0.047	0.046	0.046	0.046
$\beta = 0.4$	1	1	1	1	1	1
$\beta = 0.6$	1	1	1	1	1	1
CASE V						
$\beta = 0$	*0.135*	0.04	0.043	0.041	0.041	0.041
$\beta = 0.4$	0.994	0.971	0.974	0.971	0.971	0.971
$\beta = 0.6$	1	1	1	1	1	1
CASE VI						
$\beta = 0$	*0.135*	0.04	0.043	0.041	0.041	0.041
$\beta = 0.4$	0.986	0.939	0.948	0.939	0.939	0.939
$\beta = 0.6$	1	0.999	0.999	0.999	0.999	0.999

A Weight function is applied to adjust for the gene size.*

**The nominal error rate is set to be 0.05. Type I error rates are listed when $\beta = 0$. Power is listed when $\beta > 0$. Inflated Type I error rates are italicized.*

***A weight function $w_i = 2/\sqrt{n_i}$ is applied to each test to adjust for the size of gene.*

$\beta_1 = \ldots = \beta_6 = 0$. As shown in **Tables 1**, **2**, the Type I error rate for the original Lancaster procedure is inflated (>0.05) for all of the six cases. In contrast, five modified Lancaster procedures ($T_A - T_E$) have well controlled Type I error rates (<0.05).

The power of all test statistics was compared for regression coefficient values set at $\beta = 0.4$ and $\beta = 0.6$, respectively. The results in **Tables 1**, **2** suggest strong and comparable power among the modified Lancaster procedures. In most simulated cases, the proposed methods have more than 80% power to detect $\beta = 0.4$. When the effect size increases to $\beta = 0.6$, the power of proposed methods increases to 90% or above. Also the power of these tests improves as sample size increases from $n = 200$ to $n = 400$.

We simulate different levels of LD for SNPs with $D = 0$, 1.5, 2, and uniform(0, maximum of LD). To save the space, we only show the results for $D = 1.5$ (**Table 3**) and $D = 2$ (**Tables 1**, **2**). Our findings show that the inflation of Type I error rate for the original Lancaster procedure gets severe when LD is strong (**Tables 1**, **2**). The modified Lancaster procedures ($T_A - T_E$) have well-controlled Type I error rates and power for both moderate and strong LD (**Tables 1–3**).

In **Table 4**, we assess the performance of all tests without a weighting function. We then compare the results in **Table 4** (without a weight function) vs. **Table 1** (with a weight function). All other simulation parameters are held the same in **Tables 1**, **4**. We note that the original Lancaster procedure without a weighting function (**Table 4**) tends to have higher Type I error rates than the original Lancaster procedure with a weighting function (**Table 1**). For modified tests ($T_A - T_E$), the power is increased when a weighting function is used. This confirms that an appropriate weight function is beneficial to the Lancaster procedure.

CASE STUDY: RENAL TRANSPLANT TOLERANCE DATA

We revisited a kidney transplant data first collected and analyzed by Newell et al. (2010). Data were downloaded from the GEO website with ID = GDS4266 (http://www.ncbi.nlm.nih.

Table 3 | Type I error and power for independent Lancaster Procedure and five approximations to correlated Lancaster Procedures when sample size = 400 and linkage disequilibrium $D = 0.15$.

	Independent Lancaster procedure	T_A	T_B	T_C	T_D	T_E
CASE I						
$\beta = 0$	*0.066*	0.043	0.045	0.043	0.044	0.043
$\beta = 0.4$	0.991	0.978	0.978	0.978	0.978	0.978
$\beta = 0.6$	1	1	1	1	1	1
CASE II						
$\beta = 0$	*0.059*	0.031	0.035	0.031	0.031	0.031
$\beta = 0.4$	0.978	0.964	0.967	0.964	0.964	0.964
$\beta = 0.6$	1	1	1	1	1	1
CASE III						
$\beta = 0$	*0.053*	0.029	0.034	0.029	0.03	0.029
$\beta = 0.4$	0.898	0.836	0.844	0.837	0.837	0.836
$\beta = 0.6$	0.999	0.996	0.997	0.996	0.996	0.996
CASE IV						
$\beta = 0$	*0.072*	0.041	0.045	0.041	0.041	0.041
$\beta = 0.4$	0.977	0.962	0.964	0.962	0.962	0.962
$\beta = 0.6$	1	1	1	1	1	1
CASE V						
$\beta = 0$	*0.072*	0.041	0.045	0.041	0.041	0.041
$\beta = 0.4$	0.946	0.899	0.905	0.9	0.901	0.899
$\beta = 0.6$	0.999	0.996	0.996	0.996	0.996	0.996
CASE VI						
$\beta = 0$	*0.072*	0.041	0.045	0.041	0.041	0.041
$\beta = 0.4$	0.807	0.732	0.045	0.733	0.733	0.732
$\beta = 0.6$	0.978	0.965	0.045	0.965	0.965	0.965

A weight function is applied to adjust for the gene size.*

**The nominal error rate is set to be 0.05. Type I error rates are listed when $\beta = 0$. Power is listed when $\beta > 0$. Inflated Type I error rates are italicized.*

**A weight function $w_i = 2/\sqrt{n_i}$ is applied to each test to adjust for the size of gene.*

Table 4 | Type I error and power for independent Lancaster Procedure and five approximations to correlated Lancaster Procedures when sample size = 200 and linkage disequilibrium $D = 0.20$.

	Independent Lancaster procedure	T_A	T_B	T_C	T_D	T_E
CASE I						
$\beta = 0$	*0.106*	0.027	0.03	0.027	0.027	0.027
$\beta = 0.4$	1	0.997	0.997	0.997	0.997	0.997
$\beta = 0.6$	1	1	1	1	1	1
CASE II						
$\beta = 0$	*0.1*	0.029	0.03	0.029	0.029	0.029
$\beta = 0.4$	0.935	0.801	0.812	0.801	0.803	0.801
$\beta = 0.6$	0.998	0.976	0.98	0.976	0.977	0.976
CASE III						
$\beta = 0$	*0.118*	0.041	0.042	0.041	0.041	0.041
$\beta = 0.4$	0.608	0.307	0.32	0.307	0.307	0.307
$\beta = 0.6$	0.881	0.663	0.679	0.665	0.666	0.663
CASE IV						
$\beta = 0$	*0.115*	0.037	0.04	0.038	0.038	0.037
$\beta = 0.4$	1	0.994	0.994	0.994	0.994	0.994
$\beta = 0.6$	1	1	1	1	1	1
CASE V						
$\beta = 0$	*0.115*	0.037	0.04	0.038	0.038	0.037
$\beta = 0.4$	0.78	0.487	0.5	0.488	0.489	0.487
$\beta = 0.6$	0.977	0.869	0.882	0.869	0.87	0.869
CASE VI						
$\beta = 0$	*0.115*	0.037	0.04	0.038	0.038	0.037
$\beta = 0.4$	0.782	0.579	0.589	0.579	0.58	0.579
$\beta = 0.6$	0.964	0.885	0.888	0.885	0.885	0.885

No Weight function is applied to adjust for the gene size.*

**The nominal error rate is set to be 0.05. Type I error rates are listed when $\beta = 0$. Power is listed when $\beta > 0$. Inflated Type I error rates are italicized.*

**These are the un-weighted tests with $w_i = 2$ for all genes. We do not adjust the size of genes.*

gov/sites/GDSbrowser?acc=GDS4266). A group of tolerant renal transplant recipients (Tolerant, $n = 19$), as defined by stable graft function in the absence of immunosuppression for more than 1 year, were compared to subjects with stable graft function who were receiving standard immunotherapy (SI, $n = 27$) as well as to a group of healthy controls (Control, $n = 12$). Gene expression profiles of whole-blood total RNA from all subjects were measured by microarray. The goal of the study was to identify genetic variants associated with long-term allograft survival without the requirement for continuous immunosuppression, a condition known as allograft tolerance. Newell et al. (2010) performed statistical analysis to identify differentially expressed genes between the SI group and the Tolerant group. The results revealed a critical role for B cells in regulating alloimmunity, and provided a candidate set of genes for wider-scale screening of renal transplant recipients. However, no comprehensive pathway analysis was conducted by this group (Newell et al., 2010).

To further understand molecular mechanisms underlying renal allograft tolerance, we have applied the modified Lancaster procedure to this dataset to identify candidate cellular pathways. Gene expression levels were normalized using Robust Multichip Average (rma) preprocessing methodology, which included background subtraction, quantile normalization, and summarization via median-polish.

Gene expression levels were summarized for a total of 54,675 probes from 21,049 genes. Expression levels were compared among three groups using the Bioconductor "Limma" package. Three pair wise comparisons were conducted, including: SI vs. Control, SI vs. Tolerant, and Tolerant vs. Control. Then three comparisons were combined into one F-test. This is equivalent to a One-Way ANOVA for each gene except that the residual mean squares have been moderated across genes. P-values from multiple hypothesis testing were adjusted by FDR (Benjamini and Hochberg, 1995). Our results of differentially expressed genes are consistent with the previous published work. See Newell et al. (2010) for the gene analysis findings.

Although (Newell et al., 2010) identified a set of differentially expressed genes, our analysis demonstrates that these significant

Table 5 | Top 10 significant pathways detected by the modified Lancaster procedure (T_A).

GO accession	Pathway name	Gene ontology	URL	#Gene	#Probe	Adjusted *P*-value
GO:0030183	B cell differentiation	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/B_CELL_ DIFFERENTIATION	12	29	0.003541
GO:0042113	B cell activation	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/B_CELL_ ACTIVATION	20	45	0.003541
GO:0003823	Antigen binding	Molecular function	http://www.broadinstitute.org/gsea/msigdb/cards/ANTIGEN_ BINDING	23	51	0.003541
GO:0004709	Map kinase kinase kinase activity	Molecular function	http://www.broadinstitute.org/gsea/msigdb/cards/MAP_KINASE_KINASE_KINASE_ACTIVITY	10	32	0.003541
GO:0017148	Negative regulation of translation	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/NEGATIVE_REGULATION_OF_TRANSLATION	23	36	0.003541
GO:0042493	Response to drug	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/RESPONSE_ TO_DRUG	20	35	0.004669
GO:0001772	Immunological synapse	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/IMMUNOLOGICAL_SYNAPSE	10	18	0.006603
GO:0030098	Lymphocyte differentiation	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/LYMPHOCYTE_DIFFERENTIATION	26	53	0.007986
GO:0042036	Negative regulation of cytokine biosynthetic process	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/NEGATIVE_REGULATION_OF_CYTOKINE_BIOSYNTHETIC_PROCESS	12	21	0.008582
GO:0009890	Negative regulation of biosynthetic process	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/NEGATIVE_REGULATION_OF_BIOSYNTHETIC_PROCESS	30	48	0.008582

genes have small effect sizes with fold changes <1.5. Therefore, a limited number of individual genes in the absence of a biological context is inadequate to explain the total variation of allograft tolerance among renal transplant patients.

To address this issue, we performed the modified Lancaster procedure (T_A) as described in Section Correlated Lancaster Procedures to combine p-values from pathways. Combining p-values allows us to integrate small effects in pathway and gain the power of statistical testing. A total of 1454 Gene Ontology human pathway gene sets were analyzed. The size of pathways ranged from 9 genes to 2131 genes, with a median of 27 genes per pathway. Also, the number of probes per gene was highly variable. In order to map genes to pathways, we removed genes without gene symbols from the analysis. Among 21,049 genes with gene symbols, approximately 48% ($n = 10161$) of genes were interrogated with a single probe, 26% ($n = 5389$) of genes were queried using 2 probes, 14% ($n = 2842$) of genes were assessed by 3 probes. There were 3 or more probes for each on the remaining genes (range: 4–17). This finding indicates that larger genes would have more p-values and a stronger impact to pathway analysis. To prevent this bias, we set the weight function as $w_i = 2/\sqrt{n_i}$ where n_i is the number of probes for the ith gene.

We performed pathway analysis for the One-Way ANOVA test and three pair wise comparisons. The top 10 significant pathways based on the One-Way ANOVA test are listed in **Table 5**. The top two pathways, B cell differentiation (GO:0030183) and B cell activation (GO:0042113), confirm the signature of B cell involvement described by Newell et al. (2010). Furthermore, we identified other pathways related to B cell activation and function. These include antigen binding (GO:0003823), map kinase kinase kinase activity (GO:0004709) and lymphocyte differentiation (GO:0030098). These pathways are biologically consistent with the proposed role of B-lymphocytes in renal transplant tolerance reported by Newell et al. In contrast, when we performed the traditional Fisher's method without considering correlation structures (LD) within pathways or applying a weighting function to compensate for variability in the number of probes per gene, the result was a list of larger pathways, some containing >1000 genes, describing more general cellular processes and not specifically related to immune functions (See **Table 6**, #gene and

Table 6 | Top 10 significant pathways detected by the traditional Fisher's method.

GO accession	Pathway name	Gene ontology	URL	# Gene	# Probes	Adjusted *P*-value
GO:0005737	Cytoplasm	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/CYTOPLASM	2078	4986	0.E+00
GO:0005634	Nucleus	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/NUCLEUS	1393	3588	0.E+00
GO:0043283	Biopolymer metabolic process	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/BIOPOLYMER_METABOLIC_PROCESS	1653	4240	0.E+00
GO:0016020	Membrane	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/MEMBRANE	1954	4395	3.E−307
GO:0006139	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_METABOLIC_PROCESS	1217	3112	6.E−305
GO:0007165	Signal transduction	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/SIGNAL_TRANSDUCTION	1604	3826	1.E−296
GO:0044425	Membrane part	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/MEMBRANE_PART	1638	3670	4.E−251
GO:0019538	Protein metabolic process	Biological process	http://www.broadinstitute.org/gsea/msigdb/cards/PROTEIN_METABOLIC_PROCESS	1205	3022	2.E−245
GO:0044422	Organelle part	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/ORGANELLE_PART	1173	2934	1.E−230
GO:0044446	Intracellular organelle part	Cellular component	http://www.broadinstitute.org/gsea/msigdb/cards/INTRACELLULAR_ORGANELLE_PART	1168	2923	4.E–230

#probe). Furthermore, when comparing the SI group and the Control group, the traditional method identified 1078 significant pathways while our proposed method narrowed the list down to 64 significant pathways (adjusted p-value <0.05). The increase in number of significant pathways identified by the traditional approach is primarily due to false positive discovery, and is consistent with the inflation of Type I error rate as presented in Section Empirical Assessments. Thus, by accounting for correlation structures (LD) within pathways and the number of probes per gene, our proposed method minimized identification of larger, non-specific cellular processes pathways, and instead revealed more focused and functionally relevant biological pathways implicating a role for a humoral immune response in immunotolerance to renal transplants (See **Table 5**, #gene and #probe).

DISCUSSION AND CONCLUSIONS

Modifications to the Lancaster procedure are proposed to take correlations among p-values into account. Extensive simulation studies show that the original Lancaster procedure has inflated Type I error rates due to correlation among p-values. By using permutation approach to estimate the correlation among p-values, the proposed methods have well-controlled Type I error rates and maintain strong power to detect signals related to SNPs in pathways.

Among five proposed approximation methods ($T_A, \ldots, T_E$), the Satterthwaite approximation (T_A) is the most computationally efficient. Other approximation methods ($T_B, \ldots, T_E$) are based on the Satterthwaite approximation. Therefore, we recommend using the Satterthwaite approximation (T_A) as the standard procedure to modify the Lancaster procedure. Among other approximation methods, simulation results in Section Correlated Lancaster Procedures show that, for data with stronger internal correlation, T_D and T_E have better approximation than T_B and T_C. Our simulation study and the case study further provide evidence that T_D tends to have slightly higher power than the Satterthwaite approximation T_A. The R code for five approximation is posted at http://d.web.umkc.edu/daih/.

ACKNOWLEDGMENTS

We thank two reviewers for their constructive comments, which helped us improve the manuscript. This work was supported in part by NSF grant DMS-1209112 (to Yuehua Cui).

REFERENCES

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Stat.* 38, 303–324. doi: 10.1214/aoms/1177698949

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–833. doi: 10.2307/2346101

Benjamini, Y., and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scand. J. Stat.* 24, 407–417. doi: 10.1111/1467-9469.00072

Cheng, C., and Pounds, S. (2007). False discovery rate paradigms for statistical analyses of microarray gene expression data. *Bioinformation* 1, 436–446. doi: 10.6026/97320630001436

Cui, Y., Kang, G., Sun, K., Qian, M., Romero, R., and Fu, W. (2008). Gene-centric genomewide association study via entropy. *Genetics* 179, 637–650. doi: 10.1534/genetics.107.082370

Dai, H., Bhandary, M., Becker, M. L., Leeder, S. J., Gaedigk, R., and Motsinger-Reif, A. A. (2012a). Global tests of p-values for multifactor dimensionality reduction models in selection of optimal number of target genes. *BioData Min.* 5:3. doi: 10.1186/1756-0381-5-3

Dai, H., Charnigo, R., Srivastava, T., Talebizadeh, Z., and Ye, S. (2012b). Integrating P-values for genetic and genomic data analysis. *J. Biom. Biostat.* 3:e117. doi: 10.4172/2155-6180.1000e117

Dai, H., and Charnigo, R. (2008). Omnibus testing and gene filtration in microarray data analysis. *J. Appl. Stat.* 35, 31–47. doi: 10.1080/02664760701683528

Fisher, R. A. (1922). On the interpretation of x^2 from contingency tables and calculation of p. *J. R. Stat. Soc. A* 85, 87–94. doi: 10.2307/2340521

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Fisher, R. A., and Cornish, E. A. (1960). The percentile points of distributions having known cumulants. *Technometrics* 2, 209–225. doi: 10.1080/00401706.1960.10489895

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* 93, 509–524. doi: 10.1093/biomet/93.3.509

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280. doi: 10.1093/nar/gkh063

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375. doi: 10.1371/journal.pcbi.1002375

Koziol, J. A. (1996). A note on Lancaster's procedure for the combination of independent events. *Biom. J.* 38, 653–660. doi: 10.1002/bimj.4710380603

Lancaster, H. D. (1961). The combination of probabilities: an application of orthonomral functions. *Aust. J. Stat.* 3, 20–33. doi: 10.1111/j.1467-842X.1961.tb00058.x

Li, S., Williams, B. L., and Cui, Y. (2011). A combined p-value approach to infer pathway regulations in eQTL mapping. *Stat. Interface* 4, 389–402. doi: 10.4310/SII.2011.v4.n3.a13

Littell, R. C., and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Am. Stat. Assoc.* 66, 802–806. doi: 10.1080/01621459.1971.10482347

Littell, R. C., and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests. II. *J. Am. Stat. Assoc.* 68, 193–194. doi: 10.1080/01621459.1973.10481362

Lugannani, R., and Rice, S. O. (1980). Saddlepoint approximation for the sum of independent random variables. *Adv. Appl. Probab.* 12, 475–490. doi: 10.2307/1426607

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 37, D619–D622. doi: 10.1093/nar/gkn863

Mosteller, F., and Bush, R. R. (1954). Selected quantitative technique. *Handb. Soc. Psychol.* 1, 289–334.

Newell, K. A., Asare, A., Kirk, A. D., Gisler, T. D., Bourcier, K., Suthanthiran, M., et al. (2010). Identification of a B cell signature associated with renal transplant tolerance in humans. *J. Clin. Invest.* 120, 1836–1847. doi: 10.1172/JCI39933

Nikolsky, Y., and Bryant, J. (2009). Protein networks and pathway analysis. Preface. *Methods Mol. Biol.* 563, v–vii. doi: 10.1007/978-1-60761-175-2

Patnaik, P. B. (1949). The non-central x^2 - and F - distributions and their applications. *Biometrika* 36, 202–232.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., et al. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* 18, 111–117. doi: 10.1038/ejhg.2009.115

Stouffer, S., DeVinney, L., and Suchmen, E. (1949). *The American Solder: Adjustment during Army Life.* Vol. 1. Princeton, NJ: Princeton University Press.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* 96, 434–442. doi: 10.1093/jnci/djh075

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81, 1278–1283. doi: 10.1086/522374

Wilson, E. B., and Hilferty, M. M. (1931). The distribution of chi-square. *Proc. Natl. Acad. Sci. U.S.A.* 17, 684–688. doi: 10.1073/pnas.17.12.684

Wu, M. C., and Lin, X. (2009). Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Stat. Methods Med. Res.* 18, 577–593. doi: 10.1177/0962280209351925

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 November 2013; accepted: 27 January 2014; published online: 20 February 2014.

*Citation: Dai H, Leeder JS and Cui Y (2014) A modified generalized Fisher method for combining probabilities from dependent tests. Front. Genet. **5**:32. doi: 10.3389/fgene.2014.00032*

This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics.