Published in final edited form as:

Theor Popul Biol. 2010 February ; 77(1): 1. doi:10.1016/j.tpb.2009.10.002.

Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions

Jason Gertz^{*}, Justin P. Gerke^{*}, and Barak A. Cohen^{*,1}

^{*}Center for Genome Sciences, Department of Genetics, Washington University in St. Louis School of Medicine, 4444 Forest Park Blvd., St. Louis, MO 63108

Abstract

With technological advances in genetic mapping studies more of the genes and polymorphisms that underlie Quantitative Trait Loci (QTL) are now being identified. As the identities of these genes become known there is a growing need for an analysis framework that incorporates the molecular interactions affected by natural polymorphisms. As a step towards such a framework we present a molecular model of genetic variation in sporulation efficiency between natural isolates of the yeast, *Saccharomyces cerevisiae*. The model is based on the structure of the regulatory pathway that controls sporulation. The model captures the phenotypic variation between strains carrying different combinations of alleles at known QTL. Compared to a standard linear model the molecular model requires fewer free parameters, and has the advantage of generating quantitative hypotheses about the affinity of specific molecular interactions in different genetic backgrounds. Our analyses provide a concrete example of how the thermodynamic properties of protein-protein and protein-DNA interactions naturally give rise to epistasis, the non-linear relationship between genotype and phenotype. As more causative genes and polymorphisms underlying QTL are identified, thermodynamic analyses of quantitative traits may provide a useful framework for unraveling the complex relationship between genotype and phenotype.

Keywords

Quantitative Traits; Epistasis; Thermodynamic models; Gene regulation; Yeast

Introduction

Understanding the molecular basis of complex traits is a major challenge in modern Genetics. Faster and more efficient genotyping methods have propelled studies that seek to identify QTL in many different systems, including humans (Altshuler *et al.* 2008). The genes, and even the specific polymorphisms, that underlie QTL are now being identified. As we learn the molecular identity of causative polymorphisms, the challenge will be to understand the mechanisms by which these molecular changes generate phenotypic differences between individuals.

Despite this challenge, most methods for the analysis of quantitative traits do not incorporate information about the types of genes and alleles under study. The traditional formalism for analyzing quantitative traits was laid down by R.A. Fisher more than ninety years ago (Fisher

¹Corresponding Author BAC, Tel: 314-362-3674, Fax: 314-362-2156, cohen@genetics.wustl.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1918). The simple and powerful framework Fisher enumerated remains the basis of most quantitative trait analyses today. At the heart of the method is a model based on the linear regression of phenotype on genotype. In the linear model, nonlinear genetic phenomena, such as dominance or epistasis, are captured by higher-order terms in the regression equations (Cheverud and Routman 1995; Hansen and Wagner 2001). These terms allow non-linear behaviors to be captured by linear models, but create more complex models that require more statistical power to analyze. As a result, a common approach is to lessen or remove genetic interactions by analyzing data on transformed scales (Lynch and Walsh 1998).

Do these interaction terms reflect “real” molecular interactions, or are they just statistical corrections that are necessary for a linear model to capture the non-linear relationship between genotype and phenotype? There have been relatively few studies that have attempted to reconcile the non-linear effects of genotype on phenotype with the nonlinear biochemical processes that must underlie this relationship. Wright (Wright 1934) first argued that non-linearities in the genotype/phenotype relationship were likely a result of the inherent properties of how flux moves through metabolic networks. Kacser and Burns (Kacser and Burns 1973; Kacser and Burns 1981) proved that a non-linear relationship between genotype and phenotype will exist for most genes that function in any metabolic network. Gibson (Gibson 1996) expanded this theoretical framework to show that the thermodynamic properties of gene regulatory networks can also generate dominance and epistasis. Here, we provide an example of how specific natural genetic variants in a regulatory network of transcription factors can lead to a non-linear relationship between genotype and phenotype.

We present a molecular model of quantitative genetic variation in sporulation efficiency, a highly heritable trait in natural isolates of *Saccharomyces cerevisiae* (Gerke *et al.* 2006). The model is based on a description of the molecular interactions that govern the decision to sporulate, rather than on a statistical description of the genotype/phenotype relationship. The effects of polymorphism are modeled as changes in the parameters that govern the interactions between genes in the sporulation pathway. We show that the molecular model captures the genotype/phenotype relationship as well as the standard linear model, but with half the number of free parameters and with the advantage of generating specific testable hypotheses about the affinity of particular molecular interactions in different genetic backgrounds. We provide specific examples of how nonlinear phenomena, such as epistasis, arise naturally from the sigmoidal curves that describe the formation of protein-protein and protein-DNA complexes. We conclude that thermodynamic analyses of complex traits will be a useful addition to the suite of analytical tools now employed to dissect these traits.

Results

Yeast strains isolated from oak trees exhibit high sporulation efficiency relative to vineyard-derived isolates (Gerke *et al.* 2006). Sporulation efficiency is defined as the fraction of cells in a culture which undergo meiosis and form four-spore tetrads when the culture is starved for nitrogen and for fermentable carbon. We recently showed that natural variation in sporulation efficiency between an oak tree and vineyard strain is due largely to four single nucleotide polymorphisms (SNPs) in three transcription factors (TFs) (Gerke *et al.* 2009). *RSF1*, which encodes a transcription factor that promotes mitochondrial activity (Lu *et al.* 2003) contains a causative coding SNP in its open reading frame (ORF). *IME1*, a transcription factor whose activity promotes sporulation, contains both a causative coding SNP in its ORF and a causative non-coding SNP in its promoter. *RME1*, a transcription factor that represses the transcription of *IME1* contains a single causative SNP in its promoter. Together these four causative changes account for a large fraction of the variation in sporulation efficiency between the oak and vineyard parent strains.

Our initial analysis of these four variants with a standard linear model revealed extensive epistasis between all causative variants. All two- and three-way interactions, as well as the four-way interaction, were statistically significant (Gerke et al. 2009). Because we could place the three TFs underlying variation in sporulation into a molecular pathway, we constructed a molecular model of the effects of these causative polymorphisms, as a way to address the meaning of the epistatic terms in the linear model.

We built a molecular model of yeast sporulation centered on the causal variants (Figure 1). The structure of this model is guided by previous knowledge of the transcription factors affecting sporulation efficiency. The first section of the model is a thermodynamic model of *IME1* regulation, where Rme1 represses (Covitz and Mitchell 1993), and Rsf1 promotes the transcription rate of *IME1* (Jambhekar and Amon 2008) (see Methods). In the model the allelic status of the *RSF1* and *RME1* variants directly affect the transcription rate of *IME1*. The *IME1* non-coding region contains a causative variant in a potential Rme1 binding site. The allelic status of this variant is assumed to affect the affinity of Rme1 for the *IME1* promoter, which in turn affects the strength of *IME1* repression by *RME1*. We capture the effects of these three polymorphisms on *IME1* expression using a thermodynamic framework we previously used to analyze the effects of combinations of cis-regulatory sites on gene expression (Gertz and Cohen 2009; Gertz et al. 2009).

The fourth polymorphism, the coding change in *IME1*, is modeled as affecting the equilibrium between the active and inactive forms of the Ime1 protein. Ime1 is the master regulator of sporulation and the levels of active Ime1 in large part determine whether a cell will sporulate (Kassir et al. 1988; Nachman et al. 2007). Ime1 is bound and phosphorylated by Rim11, which activates Ime1 and drives the decision to sporulate (Bowdish et al. 1994). The coding polymorphism in *IME1* occurs at amino acid L325, in a region required for the interaction of Ime1 with Rim11 (Rubin-Bejerano et al. 1996). In the model, the equilibrium binding constant ($K_{\text{Ime1-Rim11}}$) between Ime1 and Rim11 depends on the allelic status of the *IME1* coding SNP. The relative concentration of Rim11-bound Ime1, which depends on the predicted expression level of Ime1 (above) and the affinity of Rim11 for Ime1, is captured by a Hill equation (Methods). The predicted relative concentrations of Rim11-bound Ime1 are used as predictions of the relative sporulation efficiencies between strains with different genotypes.

We directly compared the performance of the molecular model and the linear regression model. We estimated the parameters of each model using the genotypes and phenotypes of a series of allele replacement strains in which all possible combinations of vineyard alleles were transferred into the oak strain background (Table 1). We then tested the accuracy of the trained model by predicting the sporulation efficiencies of 355 genotyped recombinant segregants from a cross between the oak and vineyard parents (data previously described in (Gerke et al. 2009)). The linear regression model explains 77% of the variation in sporulation efficiency of the recombinant segregants, while the molecular model explains 80% of the variation in sporulation efficiency (Figure 2). Both models systematically under predict the sporulation efficiency of recombinant segregants. This is because the isogenic allele replacement panel used as the training set cannot capture the effects of segregating minor effect QTL in the panel of recombinants. The molecular model provides a modest gain in accuracy; however, the molecular model has some advantages over the linear model. The molecular model requires half the number of free parameters as the linear model. The linear model requires sixteen parameters to capture the main effects and all of the significant interactions between causal variants. In contrast, the molecular model requires only eight parameters, and captures slightly more of the phenotypic variation.

Another advantage of the molecular model is that the parameter values of the model represent specific molecular hypotheses. The model makes explicit predictions about the fraction of

bound species in specific protein-protein and protein-DNA complexes in different genetic backgrounds. For example, the model predicts that a strain containing the wine allele of *RME1* should have 3.5-fold more binding of Rme1 to the *IME1* promoter compared to the oak parent. A strain carrying the wine promoter variant of *IME1* is predicted to exhibit 2.4-fold more binding of Rme1. While a strain carrying both the wine allele of *RME1* and the wine promoter variant of *IME1* is predicted to have 4-fold more binding of Rme1 to the *IME1* promoter compared to the oak parent. The molecular model is as good as the linear model at capturing the genotype/phenotype relationship, but has the advantage of providing experimentally testable hypotheses.

Epistasis is the phenomenon whereby the effect of changing an allele depends on the allelic status of another locus or loci. In the linear model all possible interactions between variants were significant. The molecular model provides a natural explanation for the origin of these interaction terms. In the molecular model, the curves that relate changes in gene activity due to the causal variants to the formation of active (Rim11 bound) Ime1 are sigmoidal in shape. Additive changes in the activities of the TFs in the sporulation pathway therefore produce non-additive effects on sporulation efficiency. For example, the non-coding variants in *IME1* and *RME1* have a strong interaction term in the linear model. This “interaction” is captured in the molecular model without an additional term. Based on the properties of molecular interactions, our model predicts that the combination of these two alleles will have non-additive effects on the binding of Rme1 to the *IME1* promoter, which in turn produces non-additive effects on sporulation efficiency. The model captures the fact that the effect of switching the *IME1-NC* allele from the oak to the vineyard variant is larger in the context of the *RME1-NC* vineyard allele than the *RME1-NC* oak allele (Figure 3). This example demonstrates how the non-linear curves that describe molecular interactions naturally produce a non-linear relationship between genotype and phenotype.

Discussion

We constructed a molecular model of genetic variation in sporulation efficiency between natural isolates of yeast. The causal polymorphisms reside in three transcription factors and exert their influence on phenotype by altering the protein-protein and protein-DNA interactions that control the decision to sporulate. The molecular model captures the non-linear effects of combinations of these variants with fewer parameters than the standard linear model. We interpret this result to mean that the molecular model better accounts for the non-linear biochemical processes that underlie the formation of molecular complexes. We also show how epistasis is a natural result of the sigmoidal curves that describe how these molecular complexes form. The specific hypotheses our model generates provide a clear experimental route to either validating or falsifying the structure of our model. Testing these hypotheses will increase our understanding of the basic molecular processes that generate complex traits.

Our molecular model does not provide a complete description of every biochemical event leading to sporulation, but does capture the effects of the relevant genetic variants. An alternative to the molecular model is to apply data transformations that make the data more linear. Like our molecular model, both the logit (Gianola 1980) and probit (Dempster and Lerner 1950) transformations have proven useful in allowing non-linear data to be analyzed by reducing the number of necessary parameters. In addition, numerous other transformations can be used to “linearize” the data. These methods are valid, and have shown success in mapping the basis of phenotypes that violate the assumptions of linear models, but these methods lack an explicit biochemical basis. The parameters and assumptions of these models do not reflect the underlying genes and alleles, and do not have a specific molecular interpretation. In contrast, our molecular model of variation in sporulation efficiency provides

testable hypotheses of the molecular basis for the complex relationship between genotype and phenotype in this trait.

Since the genes harboring causative variants underlying sporulation efficiency have been placed into a general gene regulatory network, there were obvious ways to connect genes in the molecular model. In other genetic systems connections between causative genes will not be as clear. We have shown that accurate molecular-based models can be built to capture quantitative genetic variation with some prior information. Further work is needed to determine if accurate molecular models can be constructed for more ambiguous genetics systems.

Materials and Methods

Data

Strains and phenotypic measurements are those used in (Gerke et al. 2009). The sporulation efficiency of strains with varying genotypes at the four causal polymorphisms was measured using a high-throughput fluorescent reporter gene assay (Gerke et al. 2006).

Model of sporulation efficiency

All calculations were performed using Matlab from The Mathworks, Inc. (Natick, MA). The model of sporulation efficiency consists of two parts. The first part captures the relative expression level of *IME1*. To model the regulation of the *IME1* promoter, we implemented a previously described thermodynamic model of gene regulation (Gertz and Cohen 2009; Gertz et al. 2009; Shea and Ackers 1985). The model treats gene expression as the equilibrium binding of transcription factors (TFs) to DNA, to each other and to RNAP. Relative gene expression levels are then assumed to be equal to the relative probabilities of RNAP being bound to a promoter.

Calculation of the probability of RNAP being bound to a promoter requires q parameters, which describe DNA-protein interactions and are equal to the natural log of the K_d of the interaction minus the natural log of the protein concentration, and ω parameters, which describe protein-protein interactions and are equal to the change in free energy (ΔG) of the interaction. Given these parameters, a Boltzmann weight for every possible state can be calculated. Summing the Boltzmann weights for the states with RNAP bound and dividing it by the sum of all Boltzmann weights gives the probability of RNAP binding to the promoter. For the *IME1* promoter, with one Rsf1 binding site and one Rme1 binding site yields:

$$Z(\text{bound}) = e^{-q_{\text{RNAP}}} (1 + e^{-(q_{\text{Rsf1}} + \omega_{\text{Rsf1-RNAP}})} + e^{-(q_{\text{Rme1}} + \omega_{\text{Rme1-RNAP}})} + e^{-(q_{\text{Rsf1}} + \omega_{\text{Rsf1-RNAP}} + q_{\text{Rme1}} + \omega_{\text{Rme1-RNAP}} + \omega_{\text{Rsf1-Rme1}})})$$

$$Z(\text{unbound}) = 1 + e^{-q_{\text{Rsf1}}} + e^{-q_{\text{Rme1}}} + e^{-(q_{\text{Rsf1}} + q_{\text{Rme1}} + \omega_{\text{Rsf1-Rme1}})}$$

$$P(\text{RNAP bound}) = \frac{Z(\text{bound})}{Z(\text{bound}) + Z(\text{unbound})}$$

where q_X is the q value of protein X interacting with its respective binding site and ω_{X-Y} is the ω value of protein X interacting with protein Y. In the model, it is assumed that Rsf1 directly activates the expression of *IME1*. However, if Rsf1 indirectly activates the expression of

IME1 through another transcription factor the model would be mathematically equivalent. Allowing for an interaction between Rsf1 and Rme1 did not significantly improve the model.

The second portion of the model represents the binding of Ime1 to the kinase Rim11. A non-cooperative Hill equation with a Hill coefficient equal to one is used to determine the fraction of Rim11 bound by Ime1 (θ):

$$\theta(\text{Ime1 bound}) = \frac{[\text{Ime1}]}{[\text{Ime1}] + K_{\text{Ime1-Rim11}}}$$

The relative expression of *IME1* calculated in the first part of the model is used as input in the second part. It is assumed that the relative probability of RNAP bound to the *IME1* promoter is equal to the relative concentration of Ime1 protein. There is one new parameter in this portion of the model $K_{\text{Ime1-Rim11}}$, the Ime1 concentration producing half occupancy of Rim11. This parameter needs to be fit for both alleles of the *IME1* coding SNP. The relative amount of active Ime1 (bound by Rim11) is assumed to be proportional to sporulation efficiency.

Parameter Fitting

The model consists of two types of parameters, allele-specific and constant, which cannot be uniquely fit simultaneously because they depend on each other. Therefore, we fit the allele-specific parameters at all possible combinations of the constant parameters values in a reasonable range, between -5 and 5 for the ω and q values and between 0 and 5 for $K_{\text{Ime1-Rim11}}$, at intervals of 0.1. At each combination of constant parameters we fit the allele-specific parameters using the `nlinfit` function in Matlab, which uses the Gauss-Newton-Levenberg-Marquardt method (Marquardt and Leavenworth 1963) to find a least-squares parameter estimate. The best overall fit on the allele replacement strains, based on the R^2 value, was used as the final model and then tested on the recombinant segregants. Similar parameter sets had R^2 values within 1% of the overall best fit and experimental data is required to constrain the model and narrow down the optimal parameter set.

Acknowledgments

We thank Michael Brent and members of the Cohen Lab for critical readings of the manuscript. This work was supported by grants from the NIH (GM078222-01A1) and NSF (MCB-0543156).

References

- Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;322:881–888. [PubMed: 18988837]
- Bowdish KS, Yuan HE, Mitchell AP. Analysis of RIM11, a yeast protein kinase that phosphorylates the meiotic activator IME1. *Mol Cell Biol* 1994;14:7909–7919. [PubMed: 7969131]
- Cheverud JM, Routman EJ. Epistasis and Its Contribution to Genetic Variance-Components. *Genetics* 1995;139:1455–1461. [PubMed: 7768453]
- Covitz PA, Mitchell AP. Repression by the yeast meiotic inhibitor RME1. *Genes & Development* 1993;7:1598–1608. [PubMed: 8339935]
- Dempster ER, Lerner IM. Heritability of Threshold Characters. *Genetics* 1950;35:212–236. [PubMed: 17247344]
- Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinb* 1918;52:399–433.
- Gerke JP, Chen CT, Cohen BA. Natural isolates of *Saccharomyces cerevisiae* display complex genetic variation in sporulation efficiency. *Genetics* 2006;174:985–997. [PubMed: 16951083]

- Gerke JP, Lorenz K, Cohen BA. Interactions between polymorphisms in transcription factors control natural variation in yeast. *Science* 2009;323:498–501. [PubMed: 19164747]
- Gertz J, Cohen BA. Environment-specific combinatorial *cis*-regulation in synthetic promoters. *Mol Syst Biol* 2009;5:244. [PubMed: 19225457]
- Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* 2009;457:215–218. [PubMed: 19029883]
- Gianola D. Genetic evaluation of animals for traits with categorical responses. *J Anim Sci* 1980;51:1272–1276. [PubMed: 7204271]
- Gibson G. Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor Popul Biol* 1996;49:58–89. [PubMed: 8813014]
- Hansen TF, Wagner GP. Modeling genetic architecture: a multilinear theory of gene interaction. *Theor Popul Biol* 2001;59:61–86. [PubMed: 11243929]
- Jambhekar A, Amon A. Control of meiosis by respiration. *Curr Biol* 2008;18:969–975. [PubMed: 18595705]
- Kacser H, Burns JA. The control of flux. *Symp Soc Exp Biol* 1973;27:65–104. [PubMed: 4148886]
- Kacser H, Burns JA. The molecular basis of dominance. *Genetics* 1981;97:639–666. [PubMed: 7297851]
- Kassir Y, Granot D, Simchen G. IME1, a positive regulator gene of meiosis in *S. cerevisiae*. *Cell* 1988;52:853–862. [PubMed: 3280136]
- Lu L, Roberts G, Simon K, Yu J, Hudson AP. Rsf1p, a protein required for respiratory growth of *Saccharomyces cerevisiae*. *Curr Genet* 2003;43:263–272. [PubMed: 12734673]
- Lynch, M.; Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc.; Sunderland, MA: 1998.
- Marquardt D, Leavenworth B. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Numerical Analysis* 1963;10:241–256.
- Nachman I, Regev A, Ramanathan S. Dissecting timing variability in yeast meiosis. *Cell* 2007;131:544–556. [PubMed: 17981121]
- Rubin-Bejerano I, Mandel S, Robzyk K, Kassir Y. Induction of meiosis in *Saccharomyces cerevisiae* depends on conversion of the transcriptional repressor Ume6 to a positive regulator by its regulated association with the transcriptional activator Ime1. *Mol Cell Biol* 1996;16:2518–2526. [PubMed: 8628320]
- Shea MA, Ackers GK. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* 1985;181:211–230. [PubMed: 3157005]
- Wright S. Physiological and evolutionary theories of dominance. *American Naturalist* 1934;68:24–53.

Figure 1.

Molecular model of sporulation efficiency. Red X's represent causative variants. The state of the *RME1* allele affects the concentration of Rme1, the state of the *RSF1* allele affects the concentration of active Rsf1, the state of the *IME1* promoter affects the K_d of the Rme1-*IME1* promoter interaction, and the state of the *IME1* ORF allele affects the binding of Ime1 to a kinase, Rim11, which activates Ime1. q represents the natural log of the concentration of the protein divided by the K_d of the protein-DNA interaction. ω represents the change in free energy of the protein-protein interactions.

Figure 2.

Performance of the molecular model and the linear model. Predicted sporulation efficiency of the linear model (A) and the molecular model (B) versus observed sporulation efficiency in a panel of recombinant segregants from a cross between the oak and vineyard parents. The blue line represents perfect agreement between observed and predicted. The genotype for each segregant is shown for the *RME1* locus (red for wine, blue for oak), the *RSF1* locus (o for oak, Δ for wine) and the *IME1* locus (hollow for oak, filled for wine).

Figure 3.

Analysis of epistasis. The sigmoid curve relating binding affinity to percent sporulation causes the observed epistasis between the non-coding SNPs of *IME1* (*IME1-NC*) and *RME1* (*RME1-NC*). The X-axis is the q value ($\ln([Rme1]/K_{d-Rme1})$) for the interaction between Rme1 and the *IME1* promoter. The *RME1* non-coding variant affects the concentration of Rme1 and the *IME1* non-coding variant affects the affinity of the interaction between Rme1 and the *IME1* promoter. All strains in this analysis are in the oak parent background.

Table 1
Parameter values of the molecular model

Variable Parameters	Values $\pm$ 95% C.I.
[Rme1] (Wine/Oak)	15.05 $\pm$ 1.98
K _{d-Rme1} (Wine/Oak)	2.03 $\pm$ 0.1
[Rsf1]/K _{d-Rsf1} (Wine/Oak)	0.19 $\pm$ 0.03
K _{Ime1-Rim11} (Wine/Oak)	4.74 $\pm$ 0.55
Constant Parameters	Values
[RNAP]/K _{d-RNAP}	0.37
$\omega_{Rme1-RNAP}$	3.5
$\omega_{Rsf1-RNAP}$	-2.5
K _{Ime1-Rim11} (Oak)	0.1