

# Linking the heritability concept to systems dynamics

Arvbarhet og biologisk systemdynamikk

Philosophiae Doctor (PhD) Thesis

Yunpeng Wang

Department of Animal and Aquacultural Sciences

Norwegian University of Life Sciences

Ås 2013



Thesis number 2013:18

ISSN 1503-1667

ISBN 978-82-575-1120-3



## **Abstract**

The concept of heritability is rooted in the observation that relatives resemble one another more than expected by chance. Narrow-sense heritability is defined as the proportion of phenotypic variance that is attributable to additive genetic variation (i.e. where an allele substitution has the same effect irrespective of the rest of the genotype), while broad-sense heritability denotes the proportion of phenotypic variance caused by genetic variation including non-additive effects. Both concepts have been highly instrumental in evolutionary biology, production biology and biomedical research for several decades.

However, this successful instrumental use should not be equated with deep understanding of how underlying biology shapes narrow- and broad-sense heritability. Nor does it guarantee that these statistical definitions and associated methodology are optimally suited to deal with the recent floods of biological data.

Seeking a deeper understanding of the relationship between narrow- and broad-sense heritability in terms of biological mechanisms, I simulated genetic variation in dynamic models of biological systems. A striking result was that the ratio between narrow-sense and broad-sense heritability depended strongly on the type of regulatory architecture involved.

Applying the same approach to an ensemble of gene regulatory network models, I showed that monotonicity features of genotype-to-phenotype maps reveal deep connections between molecular regulatory architecture and heritability aspects; connections that do not materialize from the classical distinction between additive, dominant and epistatic gene actions.

Lastly, I addressed why genome-wide association studies (GWAS) have failed to identify much of the genetic variation underlying highly heritable traits. By linking computational physiology to GWAS, one can do GWAS on lower-level phenotypes that are mathematically related to each other through a dynamic model. This allows much more precise identification of the causal genetic variation, coupled with understanding of its function.

## Sammendrag

Begrepet *arvbarhet* gjenspeiler det faktum at slektninger jevnt over ligner mer på hverandre enn på andre individer. Arvbarhet i *smal forstand* defineres som andelen av fenotypisk varians som kan tilskrives additive effekter av genetisk variasjon (altså der en allel-substitusjon har samme effekt uavhengig av resten av genotypen), mens arvbarhet i *vid forstand* betegner den samlede andelen som skyldes både additive og ikke-additive effekter. Begge begrepene har vist seg nyttige i evolusjonsbiologi, produksjonsbiologi og biomedisinsk forskning over flere tiår.

Denne nytten som verktøy er imidlertid ikke ensbetydende med dyp innsikt i hvordan de to typene av arvbarhet formes av underliggende biologi. Det er heller ikke selvsagt at disse statistisk baserte definisjonene og metodene vil være de beste til å møte dagens flom av nye biologiske data.

I mitt doktorgradsarbeid har jeg belyst hvordan forholdet mellom arvbarhet i smal og vid forstand henger sammen med biologiske mekanismer, gjennom å simulere genetisk variasjon i dynamiske modeller av fysiologiske systemer. Et slående resultat var at den regulatoriske arkitekturen til systemet har mye å si for forholdstallet mellom arvbarhet i smal og vid forstand.

På lignende vis studerte jeg arvbarhet i et knippe modeller av genregulatoriske nettverk med ulike grader av monotonitet i den matematiske sammenhengen mellom genotype og fenotype. Dette avdekket dype bånd mellom arvbarhetsmønstre og molekylær regulatorisk arkitektur; sammenhenger som ikke er åpenbare ut fra det klassiske skillet mellom additive, dominante og epistatiske gen-effekter.

Til sist tok jeg for meg svakheter ved dagens statistiske metoder for å forklare hvordan variasjon i sterkt arvbare trekk styres av genetiske forskjeller mellom individer. Såkalte hel-genom-assosiasjons-studier (genome-wide association studies, GWAS) påviser ofte en mengde relevante loci med genetisk variasjon, men disse forklarer likevel bare en liten del av den observerte arvbarheten i overordnede trekk som f.eks. kroppshøyde eller sjukdomsforekomst. En mer lovende tilnærming er å koble matematisk fysiologi til GWAS. Jeg viser at man ved å gjøre GWAS på lavnivå-fenotyper som er matematisk forbundet gjennom en dynamisk modell, kan identifisere den

årsaksbestemmende genetiske variasjonen langt mer presist og samtidig øke forståelsen av dennes funksjon.

## **Acknowledgements**

I offer my sincerest gratitude to my main supervisor, Prof. Stig Omholt, and my co-supervisors Dr. Arne B. Gjuvsland and Dr. Jon Olav Vik, for their whole-hearted support throughout the PhD program. I would also like to thank all my past and current colleagues at Center of Integrative Genetics (CIGENE) and Department of Mathematical Sciences and Technology (IMT), Marte Sodeland, Øyvind Nordbø, Jeevan Karloss, Erik Plahte, Geir Halnes, Gaute T. Enevoll, Ivar Østby, Eivind Norheim and Mohammad Youssaf, for providing an enjoyable working environment these years and for valuable advices. In particular, I am very grateful to Barbara Eriksen for helping me with all the paper work that a foreign student is exposed to when he embarks on a Norwegian PhD program. I also thank the Department of Animal and Aquacultural Sciences for funding this high-risk project at the borderline between conceptual analysis, genetics and systems dynamics.

My dear friends, Duoji, Xie Li, Yoyo, BinBin, Gao hong, Qu Zhi, Bujie, and Yu Jie, thank you for all your support during my stay here in Norway.

Finally, I would like to show my deepest gratitude to my parents (baba, Wang Jianming and mama Guo Fengqing), and to my wife (Wen Li) for their endless support.

## Table of Contents

Abstract .....	3
Acknowledgements .....	6
List of Papers .....	8
Introduction.....	9
1. The genotype-phenotype map .....	9
2. The cGP research program .....	12
2.1 Systems biological models .....	12
2.2 The cGP modeling.....	12
2.3 The state of the art of cGP studies .....	14
3. Heritability and variance components .....	16
3.1 The Conceptual background .....	16
3.2 Understanding the relationship between $h^2$ and $H^2$ .....	19
3.3 The “missing heritability” problem .....	20
4. Concluding remarks .....	22
References .....	24

## List of Papers

Wang Y, Vik JO, Omholt SW, Gjuvslund AB (2012) *On the relationship between heritability and regulatory architecture*. Submitted to PLoS Comput. Biol. (10/08/2012)

Gjuvslund AB, Wang Y, Plahte E, Omholt SW (2012) *Monotonicity is a key measure of genotype-phenotype maps*. Submitted to Molecular Systems Biology (04/01/2013)

Wang Y, Gjuvslund AB, Vik JO, Smith NP, Hunter PJ, Omholt SW (2012) *Parameters in Dynamic Models of Complex Traits are Containers of Missing Heritability*. PLoS Comput. Biol. 8(4): e1002459.  
Doi:10.1371/journal.pcbi.1002459



## Introduction

*“... unless and until we uncover the ‘rules of transformation’ that connect ‘genotype space’ with ‘phenotype space’ then we cannot seriously entertain, or be satisfied with, a gene-based theory of evolution. How an individual phenotype emerges and reproduces from a given unique set of genes inherited from its sexual parents is the central question of evolutionary theory: all the rest is subsidiary”*

Gabriel Dover (Dover, 2000)

### 1. The genotype-phenotype map

The science of genetics deals with heredity and the variation of organisms (Gove and Merriam-Webster Inc., 2012). An individual's *genotype* denotes its hereditary material, or a relevant portion of it, whereas its *phenotype* denotes its observable traits of interest. Phenotypic traits include any morphological, developmental, biochemical or physiological property all the way down to the subcellular level, as well as any behavior and product of behavior (Omholt, 2012). The mathematical abstraction of the genotype-phenotype relation as a genotype-phenotype map, assigning a phenotypic outcome to each possible genotype, facilitates the characterization and comparison of the genotype-phenotype relation for different biological systems under different environmental conditions (Vik et al., 2012)

Over the last decade, as genomics and other -omics technologies have matured (Joyce and Palsson, 2006), a flood of genotype and phenotype data has become available. Full genomes have been sequenced for many species, including human, mouse, dog, worm, fly and others (Pagani et al., 2012). Phenome data include the organismal, tissue or whole cell transcript (transcriptomics), protein (proteomics), metabolite (metabolomics) and other measurements (Joyce and Palsson, 2006).

However, this wealth of information has proved difficult to incorporate in classical population and quantitative genetics. Population genetics describes the dynamics of gene frequencies due to natural selection, genetic drift, mutation and gene flow (Hedrick, 2011), thus operating mostly in genotype space. The genotype-phenotype relation is simplified to a "fitness function", which simply assigns fitness values to different genotypes. Quantitative genetics, on the other hand, focuses on changes in phenotype distributions (Falconer and Mackay, 1996; Lynch and Walsh, 1998), primarily of continuously varying traits. Even though genes and genotypes are ingredients of quantitative genetic theory, all the genotype-level concepts are expressed in terms of the means and variances of phenotypic values (Lynch and Walsh, 1998), operating primarily in phenotype space. The genotype-phenotype relation is approximated by linear regression models of phenotypic values as a function of gene content.

A more mechanistic view of the genotype-phenotype map has emerged in developmental genetics (Johnson and Porter, 2000). The mathematical theory of dynamical systems is applied by viewing physiological state as variables and parameters as proxies of genotypes (Johnson and Porter, 2000; Omholt et al., 2000; Jaeger et al., 2012). This highlights the roles that genes play in various development and physiological processes and how gene products coordinate to produce cellular phenotypes.

More generally, *systems genetics* aims to understand the integration, coordination and transmission of genetic information through molecular, cellular, and physiological networks to generate the emergent properties of the biological systems (Nadeau and Dudley, 2011). For instance, gene-mapping methodologies have been applied on transcription level by taking transcript abundance as phenotypes (Jansen and Nap, 2001; Rockman and Kruglyak, 2006; Cheung and Spielman, 2009; Cookson et al., 2009). The associated genome regions are termed expression quantitative trait loci (eQTLs). Combining information about eQTLs, differential expression patterns between individuals and the co-expression networks, not only enriches the list of putative causal genes but also sheds some light on the generating pathways of the focal phenotypes (Ayroles et al., 2009; Flint and Mackay, 2009; Swami, 2009; Capobianco, 2012). Other types of networks, i.e. protein-protein interaction

(Rual et al., 2005; Stelzl et al., 2005) and metabolites interaction (Jeong et al., 2000; Duarte et al., 2007), have also been used to prioritize candidate genes, to interpret gene-mapping results in context (Köhler et al., 2008; Zanzoni et al., 2009; Barabási et al., 2011).

A mature genetic theory will treat both genotypes and phenotypes as state variables and study the transformations between and within the state variables (Lewontin, 1974). New genotypes are formed by recombination, whereas mating, migration and natural selection operate in phenotype space. Genotype-phenotype maps represent the transformation from genotypes into phenotypes. Population and quantitative genetics have generally neglected the complexity of the genotype-phenotype maps but have developed a tremendous amount of knowledge in terms of genotype and phenotype spaces, separately. Developmental genetics and systems genetic are promising approaches to follow albeit not as mature as population and quantitative genetics. In the next section, I describe a research program framework taking into account both the static and dynamic interactions among systemic components (Rajasingh et al., 2008; Houle et al., 2010; Vik, 2011; Omholt, 2012; Wang et al., 2012).

## **2. The research program of causally cohesive genotype-phenotype modeling**

### **2.1 Systems biological models**

Mathematical modeling of biological systems dates back almost 100 years ago. The Lotka-Volterra predator-prey model (Lotka, 1920; Volterra, 1926) and the Hodgkin-Huxley action potential model (Hodgkin and Huxley, 1952) are arguably the earliest successful examples in ecology and cell biology, respectively. The accumulation of high-throughput data at multiple biological levels in past decades has put mathematical modeling at the fore of biology as the means to understand the data and the underlying mechanisms. Subcellular, cellular, tissue, organ and organism level biological systems are represented by mathematical equations (discrete or continuous, deterministic or stochastic), obeying the inherent physical-/chemical-laws (Hunter and Borg, 2003; Kohl and Noble, 2009).

Models are increasingly tested, curated and deposited in public databases for reuse. For instance, the BioModels database focuses on subcellular or cellular models, mainly about biochemical processes (Li et al., 2010), whereas the CellML repository aims for integration across multi-scale and multi-physics and including models on cellular, tissue, organ and organism levels (Lloyd et al., 2008). However, these models do not explicitly link to the realm of genetics. In the following, I describe an approach for embedding systems biology models in a setting that maintains a relation to genetic variation.

### **2.2 Causally cohesive genotype-phenotype (cGP) modeling**

Causally cohesive genotype-phenotype (cGP) modeling (Rajasingh et al., 2008; Houle et al., 2010; Omholt, 2012) denotes an approach where (1) model elements, including state variables and parameters, are associated with genes; (2) genotype variations is represented by variation in a set of parameters; (3) the model describes how phenotypes emerge from low level of processes in a causally cohesive way. Parameters are any quantities that are constant on the time scale of a model instance. An important part of cGP modeling is to view such parameters as low-level phenotypes (Omholt, 2012). In contrast to

standard quantitative or population genetic models where phenotypic values are assigned directly to genotypes, cGP models stick genotypes and phenotypes causally together through regulatory principles and mechanisms.

By studying cGP models in virtual populations a whole range of genetic phenomena can be addressed. Figure 1 outlines the cGP modeling framework. The linkage map governs the dynamics of haplotypes cross generation. The cGP model, illustrated here by a simple three-gene regulatory network, bridges the gap between genotypes and phenotypes. Four transformations are incorporated: selection, operating in the phenotype space; meiosis, producing genetic variation by sampling and recombination processes, and operating in the genotype space; mating, generating new genotypes; and finally, the genotypic variation feeds into cGP models via model parameters and manifested by phenotypic variation. In this framework, arbitrary complex cGP models and different types of linkage maps can be studied.

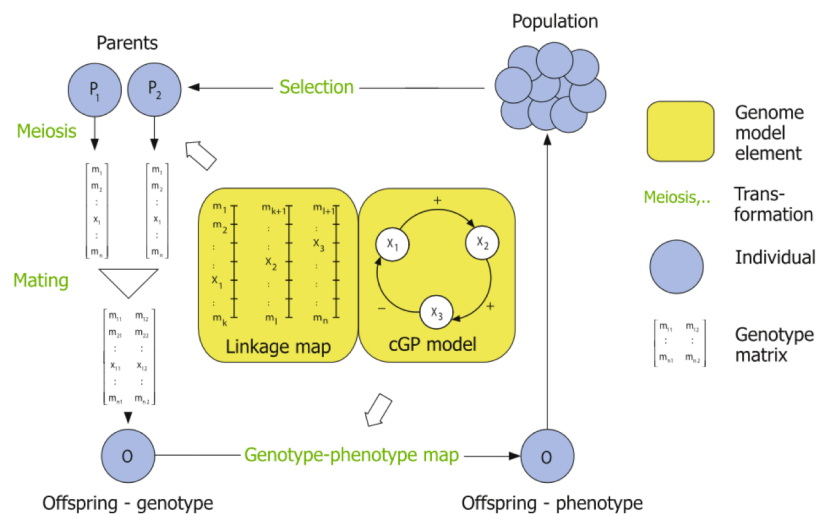


Figure 1. The scheme of cGP study framework (Omholt, 2012).

The cGPtoolbox (available at <https://github.com/jonovik/cgptoolbox>) is a Python implementation of the cGP modeling framework. It provides a streamlined interface from genomic databases to model parameter structures, and from parameters to model phenotypes. Currently, the genotype data of HapMap (Gibbs et al., 2003) populations is integrated into the pipeline and served as an example of using other data resources. Models deposited in both BioModel (Li

et al., 2010) and CellML repository (Lloyd et al., 2008) can be readily incorporated as cGP models. The simulated datasets can be analyzed using standard quantitative genetic methods. For instance, in Paper III I performed genome-wide association studies on datasets generated by the cGPtoolbox.

### **2.3 The state of the art in cGP modeling**

While the cGP modeling framework was proposed recently (Rajasingh et al., 2008; Houle et al., 2010; Vik, 2011; Omholt, 2012; Wang et al., 2012) the basic idea is not new. Jim Burns (1970) stated that: “ it is the quantitative phenotype, arising from the genotypic prescriptions and the environment, which is of critical importance for the cell’s survival and which therefore features in population genetic theory. A study of this synthetic problem would thus, by providing genotype-phenotype mappings for simple synthetic systems, help to connect two major areas of biological theory: the biochemical and the population genetics.” In this section I will illustrate the fruitfulness of the cGP modeling program from a number of early studies.

Genetic *dominance* is a phenomenon where the phenotype of the heterozygote does not fall midway between the homozygotes. The dominance concept dates back to Mendel, and was the subject of intense debates between the founders of the modern synthesis. However, geneticists paid little attention to biological mechanisms before Kacser and Burns in 1981 proposed a novel explanation based on a model of enzyme reaction kinetics (Kacser and Burns, 1981). Later studies have followed up on this. For the generic diffusion-gradient-threshold model of pattern formation, Gilchrist and Nijhout showed that this nonlinear process is capable of generating dominance in all its components (Gilchrist and Nijhout, 2001). Furthermore, in regulatory biology, transcriptional regulatory networks (Mestl et al., 1995) have the capacity of accounting for the observed dominance equally well (Omholt et al., 2000).

Interactions between genes on different loci, termed epistasis in genetics, have also been given mechanistic explanations in the last decades. Two distinct meanings are implicated by epistasis depending on the context: the physiological epistasis (Cheverud and Routman, 1995), also called functional (Hansen and Günter, 2001) or biological epistasis (Moore, 2002), describes the phenomenon that the effects of a set of genes depend on their genetic

background and, the statistical epistasis defined as the average deviation of combinations of allele effects from additivity in quantitative genetic models (Phillips, 1998; Moore, 2002; Zeng et al., 2005; Alvarez-Castro and Carlborg, 2006; Phillips, 2008). Epistasis in the biological sense is an intrinsic property of genotype-phenotype maps, but it does not necessarily turn up as statistical epistasis in given population (Cheverud and Routman, 1995). Both types of epistasis are argued to be pervasive on a range of phenotypes and across species (Cordell, 2002; Moore, 2002; Phillips, 2008; Breen et al., 2012). It is well recognized that genes or gene products function in a cooperative manner and thus physiological epistasis is expected. But how these molecular interactions give rise to the statistical epistasis is largely unknown (Moore and Williams, 2005). Using three-gene regulatory networks as cGP models, Gjuvsland et al (2007b) reported that statistical epistasis is an emergent property of functional dependence between genes. In particular, positive feedback architecture gives more pronounced statistical epistasis than other types of architectures. In a similar fashion, Pumir & Shraiman showed that additivity between loci is the norm under the regime of small parameter perturbations while with large perturbations statistical epistasis appears generally at lower order (on pair-wise level) if at all, by analyzing a validated signaling transduction model (Pumir and Shraiman, 2011). Combining experiments and mathematical modeling, Gertz et al. demonstrated that thermodynamic properties of yeast sporulation process can generate statistical epistasis phenomena (Gertz et al., 2010).

The cGP framework has also been employed to explain other genetic concepts, such as variations in phenotypic penetrance (Plahte et al., 1998; Gjuvsland et al., 2007a), the dependence of genetic variance on the shape of gene regulatory functions (Gjuvsland et al., 2007c) and genetic background (Vik et al., 2011), and the response to selection (Peccoud et al., 2004).

These examples highlight how cGP modeling goes beyond classical genetics, critically examining the mechanistic underpinnings of phenomenological descriptors used in statistical genetic research.

### **3. Heritability and variance components**

“They [offspring] resemble their parents more than remoter ancestors, and resemble those ancestors more than any chance individual”

Aristotle, c. 340 BC (Peck, 1948)

#### **3.1 Conceptual background**

Heritability is one of the most important parameters in genetics research. In evolutionary biology, knowledge of heritability of a trait is the necessary condition in predicting its response to natural selection (Visscher et al., 2008). In animal/plant breeding, it predicts the short-term response of selection operations (Falconer and Mackay, 1996; Lynch and Walsh, 1998). In gene mapping studies, it affects the efficacy of the effort in detecting causal variants (Visscher et al., 2008). In medical genetics, it affects the possibility of predicting genetic risk of diseases (Falconer and Mackay, 1996; Visscher et al., 2008; Manolio et al., 2009; Zaitlen and Kraft, 2012). However, the term heritability has been defined in several different ways. Jacquard (1983) outlines three types of definitions, biometric heritability, broad sense and narrow sense heritability (Jacquard, 1983). There also a less used meaning, realized heritability (Falconer and Mackay, 1996), which appear mainly in the artificial selection literature.

#### **Biometric heritability**

The heritability of a phenotype, here, is defined as the coefficient of the regression of offspring phenotype on mid-parent phenotype values (Jacquard, 1983). The only assumption here is that there exists a linear trend between offspring and parent phenotypes. In other words, the conditional distribution of offspring phenotype values can be expressed as a linear function of parent phenotype values. This definition only refers to the phenotype space, is a



statistical parameter that quantifies the degree of resemblance between offspring and parents.

Even though the definition of biometric heritability already indicates the method of estimation cautions need to be taken. A linear relationship between the phenotypic value of offspring and a parent (or the average of both parents) is guaranteed if the phenotype values have a bivariate normal distribution. For many phenotypes distributions close to bivariate normal are observed, but frequently nonlinear relationship is reported (Gifford and Barker, 1991; Koerhuis, 1996). If the nonlinearities are strong, transformations such as the box-cox procedure (Box and Cox, 1964) have to be done before estimation.

### **Broad sense and narrow sense genetic heritability**

The definitions of broad and narrow sense heritabilities refer to both genotypes and phenotypes. In quantitative genetics, the phenotype value,  $P$ , of an individual is assumed to be the result of a genetic effect ( $G$ ), an environmental effect ( $E$ ), and a genetic-environmental interaction effect. The combination of different factors is often assumed to be additive and the interaction term is neglected, expressed as  $P = G + E$ . The genetic contribution is further decomposed into components from additive gene effect ( $A$ ) and non-additive interactions between genes ( $D$  for dominance effect and  $I$  for epistasis effect). Accordingly, the observed phenotypic variance ( $\sigma_P^2$ ) can be decomposed into variances due to each contributing factor, i.e.,  $\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2$ , and  $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$  is the total genetic variance. Heritability in the broad sense ( $H^2$ ) is defined as the proportion of phenotypic variance that is genetic, i.e.,  $H^2 = \sigma_G^2 / \sigma_P^2$ , and, heritability in the narrow sense ( $h^2$ ) is the proportion of phenotypic variance due to additive genetic effect,  $h^2 = \sigma_A^2 / \sigma_P^2$ .

Twin study is a standard approach for estimating  $H^2$ . Monozygotic twins ( $MZ$ ) share identical genotypes and dizygotic twins ( $DZ$ ) on average share only half of their alleles (Lynch and Walsh, 1998). Twice the difference in correlation coefficients of the two types of twins is used as a estimate of  $H^2$ , i.e.,  $H^2 = 2[r(MZ) - r(DZ)]$ , under the assumption that environmental factors

contribute equally in each type (Falconer and Mackay, 1996; Lynch and Walsh, 1998).

Heritability in the narrow sense ( $h^2$ ) is traditionally estimated by variance-component framework. The covariance  $\sigma(X,Y)$ , of phenotypes of pairs of relatives ( $X,Y$ ), all with the same relationship, is expressed in terms of variances due to additive and interacting gene actions,  $\sigma_{A^m D^m}^2$  under the assumption of independent segregation and assortment principles and assuming environments do not contribute (Lynch and Walsh, 1998). *Equ.1* is the general formula for estimating genetic variance components using pedigree data (Lynch and Walsh, 1998). The summation runs over  $n$ , the number of additive effects, and  $m$ , the number of dominance effects. The exact ranges  $n$  and  $m$  depend on the relationship in pedigree. The kinship and fraternity coefficients are denoted by  $\Theta_{XY}$  and  $\Delta_{XY}$ , respectively. In most case, only additive genetic effect are interested and the genetic coefficients in *Equ.1* are determined by the kinship coefficient.

$$\sigma(X,Y) = \sum (2\Theta_{XY})^n \Delta_{XY}^m \sigma_{A^m D^m}^2 \quad (\text{Equ.1})$$

In practice, the restricted maximum likelihood (REML) algorithm is used to estimate each component. Heritability in narrow sense ( $h^2$ ) is estimated by  $h^2 = V_A / V_P$ , where  $V_A$  and  $V_P$  are the sample estimate of  $\sigma_A^2$  and  $\sigma_P^2$ , respectively.

In contrast to biometric heritability, the contributing factors to phenotypic variation are explicitly sought in genetic heritability. Thus, more assumptions have to be made both in their definitions and estimations (Jacquard, 1983). The biometric heritability is equivalent to  $h^2$  when genetic interactions involving additive effect are absent, since dominance effects does not contribute to parent-offspring resemblance. But the interpretation is different. Biometric heritability measures the degree of resemblance between parents and offspring and the reliability of using phenotypes of parents to predict offspring phenotype. Whereas,  $h^2$  measures the contributions from additive genetic variation to phenotypic variation and affect the predictability of phenotypic variation from genotypic variation and the efficiency of gene-mapping efforts (Visscher et al., 2008).

### 3.2 Understanding the relationship between $h^2$ and $H^2$

The ratio between narrow- and broad-sense heritability,  $h^2 / H^2 = \sigma_A^2 / \sigma_G^2$ , is of importance in evolutionary theory, and also has practical consequences in medical genetics and breeding. It has been intensely debated during the development of quantitative genetics, and a key question is if biochemical interactions leading to physiological dominance and epistasis will also lead to low values of this ratio. Hill et al. argued that high values of this ratio are expected, despite strongly non-linear GP maps, as long as the frequencies of causative variants are close to zero or one (Hill et al., 2008). However, line-cross populations, having intermediate gene frequencies, can also show high additive variance. In the first paper I studied five dynamic models (of the cAMP pathway, the glycolysis, the circadian rhythms, the cell cycle, and action potential cell), assuming genetic variation in model parameters. Even when imposing purely linear genotype to parameter maps and no environmental variation, we observed quite low  $V_A / V_G$  (estimate of  $\sigma_A^2 / \sigma_G^2$ ) ratios. Furthermore, models with positive feedback and cyclic dynamics gave much lower  $V_A / V_G$  ratios than those without. The results show that some types of regulatory architectures consistently maintain a transparent genotype-phenotype relation, whereas other architectures generate more subtle patterns.

This raises several issues, which are pursued in paper II. Does a positive feedback structure by itself produce low additive variance? Are there any other types of architectures that have the same capability? Why is variance so often chiefly additive, despite the ubiquitous nonlinear interactions in regulatory systems? Gjuvslund et al. proposed that high  $V_A / V_G$  ratios could be explained by monotonic relationships between the phenotypic values and gene content (i.e. the number of alleles of a given type), so-called monotonic GP maps (Gjuvslund et al., 2011). To this end, we studied an ensemble of three-gene regulatory networks, which provide the building blocks of more complex models. Moreover, we present two measures of the monotonicity of a GP map, one based on allele substitution effects, and the other based on isotonic regression (De Leeuw et al., 2009). We confirm, both numerically and analytically, that generally GP maps are indeed highly monotonic across network types.

However, regulatory motifs involving incoherent feedforward and positive feedback, as well as pleiotropy in the mapping between genotypes and gene regulatory parameters, are clearly predisposed for generating non-monotonicity. These deep connections between molecular regulatory architecture and properties of the GP maps do not materialize from the classical distinction between linear and nonlinear gene action.

### 3.3 The “missing heritability” problem

Genome-wide association studies (GWAS) find statistical associations between a complex trait and genomic variants (mostly SNPs) based on populations of unrelated individuals (Wang et al., 2005; Hardy and Singleton, 2009). But even though thousands of SNPs (<http://www.genome.gov/gwastudies/>) have been linked to complex human diseases or traits, their combined effect typically explain very little (<10%) of the heritability of a trait. This discrepancy is known as the “missing heritability” problem (Maher, 2008).

For a continuous trait GWAS is typically done as a series of univariate regressions to identify loci with significant additive effect on the trait (Zaitlen and Kraft, 2012). The proportion of phenotypic variance explained by genotypes at multiple loci ( $h_{GWAS}^2$ ) can then be estimated as follows. The phenotypic value vector ( $Y$ ) is expressed as the sum of genetic and environmental contributions,  $Y = \mu + G \cdot \alpha + \varepsilon$ ,  $Y$  contains normalized phenotype values with variance one,  $\mu$  is the vector of population mean,  $G$  is the matrix of normalized genotypes with element for  $j$ th individual at  $i$ th SNP computed by  $g'_{ij} = \frac{g_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$ ,  $g_{ij} = 0, 1, 2$  is the number of copies of the reference allele and  $p_i$  is the population frequency of the same allele,  $\alpha$  is the vector of allelic effects and the  $\varepsilon$  is the vector of environmental contributions whose elements are *i.i.d.* normal variables with mean 0 and variance  $\sigma_e^2$  (Zaitlen and Kraft, 2012). Then the additive variance  $V_A$  is computed as the sum of squared allelic effects and  $h_{GWAS}^2$  is calculated as the ratio of additive variance to total phenotypic variance. Another method used very often in GWAS is to estimate the genetic coefficient in *Equ.1* by genotypes at all genotyped SNPs and then the variance-component machinery is used to estimate each variance component (Zaitlen and Kraft, 2012).

Many possible reasons for the “missing heritability” and strategies of how to account for it have been proposed (Manolio et al., 2009; Eichler et al., 2010; Zuk et al., 2012). Estimators of heritability are based on pedigree information and phenotype and do not refer to the number of causal variants. And, depending on the type of pedigrees available, some components of gene interactions are not always separable from additive variance. On the other hand, the GWAS estimator relies on the number of causal variants, or proxies of causal variants, detected and their frequencies. So, undetected causal variants, such as SNPs with low frequencies and structure variants, incomplete linkage between common SNP and causative variants, epistatic interactions, gene-environment interaction, parent of origins, and over-estimation of  $h^2$  could all contribute to the “missing heritability” (McCarroll and Altshuler, 2007; Manolio et al., 2009; Eichler et al., 2010; Yang et al., 2010; Zuk et al., 2012).

In paper III we used cGP modeling to suggest a possible direction to tackle the problem. We integrated the HapMap III (Gibbs et al., 2003; Altshuler et al., 2010) population data into the cGP modeling framework and used the open source program simuPOP (Peng and Amos, 2010) to construct an *in silico* population suitable for performing GWAS. The action potential model was used as cGP model with 34 parameters harboring genetic variation and 16 cellular phenotypes. We showed that genome-wide association studies on parameters reveal much more genetic variation than when using higher-level cellular phenotypes. Thereby more phenotypic variation can be explained with detected SNPs. The results suggest that letting such studies be guided by computational physiology may facilitate a causal understanding of the genotype-to-phenotype map of complex traits, with strong implications for the development of phenomics technology (Houle et al., 2010).

#### **4. Concluding remarks**

The GP map concept applies to any time point in the ontogeny of a living system and it is an abstraction of a relation that is the outcome of a very complex dynamics. This is the major rationale underlying efforts to elucidate genetic concepts and understand genetic phenomena by use of systems dynamics. In this thesis I have shown that this approach does indeed reveal new understanding on matters related to the heritability concept. I found that the ratio between narrow sense and broad sense heritability of a trait is a function of the type of regulatory architectures that generate the trait. In particular, the positive feedback and incoherent feedforward loops are more prone to give low ratios than other structures due to more non-monotone genotype-phenotype map. Thus, the concept of monotonicity of genotype-to-phenotype maps reveal deeper connections between molecular regulatory architecture and heritability than the classical distinction between additive, dominant and epistatic gene actions. Finally, I demonstrated that the “missing heritability” problem arising from genome wide association studies (GWAS) could be solved partially by linking computational physiology to GWAS. In such a setting one can do GWAS on lower-level phenotypes mathematically related to each other through a dynamic model, and, more genetic variants could be discovered thus increasing the proportion of heritability accounted.

Investigation of the GP map associated with high-level phenotypes manifested at the level of the whole organism requires computational models integrating molecular-, cellular-, tissue-, and organ-level processes to high-level function. There is in principle no limit to the complexity of biological models that can be used in a cGP context. In the not too distant future, the cGP program in a multiscale and multiphysics context will probably give us an extensive understanding of how different types of genetic variation propagate and manifest in different physiological settings and genetic backgrounds.

The results of this thesis at least suggest that this emerging understanding will become a rich source for filling current genetic concepts with causal content and for identifying basic theoretical principles concerning the relationship between genetic variation, regulatory anatomy and phenotypic

variation that will provide a foundation for a quantitative genetics theory based on biological mechanism.

## References

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* *467*, 52–58.

Alvarez-Castro, J.M., and Carlborg, O. (2006). A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. *Genetics* *176*, 1151–1167.

Ayroles, J.F., Carbone, M.A., Stone, E.A., Jordan, K.W., Lyman, R.F., Magwire, M.M., Rollmann, S.M., Duncan, L.H., Lawrence, F., ANHOLT, R.R.H., et al. (2009). Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genetics* *41*, 299–307.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* *12*, 56–68.

Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations. *Journal of The Royal Statistical Society Series B (Methodological)* *26*, 211–252.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. *Nature* *490*, 535–538.

Burns, J. (1970). The synthetic problem and the genotype-phenotype relation in cellular metabolism. In: Waddington, C.H. (E.D.), *Towards a Theoretical Biology. Drafts. an I.U.B.S. Symposium.* *3*, 47–51.

Capobianco, E. (2012). Dynamic networks in systems medicine. *Frontiers in Genetics* *3*, 185.

Cheung, V.G., and Spielman, R.S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews Genetics* *10*, 595–604.

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

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* *10*, 184–194.

Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* *11*, 2463–2468.

De Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software* *32*.

Dover, G. (2000). How genomic and developmental dynamics affect evolutionary processes. *Bioessays* *22*, 1153–1159.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of The United States of America* *104*, 1777–1782.



- Eichler, E.E., Flint, J., Gibson, G., and Kong, A. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* 11, 446–450.
- Falconer, D.S., and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*. Harlow: Longman Group.
- Flint, J., and Mackay, T.F.C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* 19, 723–733.
- Gertz, J., Gerke, J.P., and Cohen, B.A. (2010). Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions. *Theoretical Population Biology* 77, 1–5.
- Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H., Ch'ang, L-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international HapMap project. *Nature* 426, 789–796.
- Gifford, D.R., and Barker, J. (1991). The nonlinearity of offspring-parent regression for total sternopleural bristle number of *Drosophila melanogaster*. *Theoretical and Applied Genetics* 82, 217–220.
- Gilchrist, M.A., and Nijhout, H.F. (2001). Nonlinear developmental processes as sources of dominance. *Genetics* 159, 423–432.
- Gjuvsland, A.B., Hayes, B.J., Meuwissen, T.H., Plahte, E., and Omholt, S.W. (2007a). Nonlinear regulation enhances the phenotypic expression of trans-acting genetic polymorphisms. *BMC Systems Biology* 1, 32.
- Gjuvsland, A.B., Hayes, B.J., Omholt, S.W., and Carlborg, O. (2007b). Statistical Epistasis Is a Generic Feature of Gene Regulatory Networks. *Genetics* 175, 411–420.
- Gjuvsland, A.B., Plahte, E., and Omholt, S.W. (2007c). Threshold-dominated regulation hides genetic variation in gene expression networks. *BMC Systems Biology* 1, 57.
- 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. *Journal of Evolutionary Biology* 24, 2269–2279.
- Hansen, T.F., and Wagner, G.P. (2001). Modeling Genetic Architecture: A Multilinear Theory of Gene Interaction. *Theoretical Population Biology* 59, 61–86.
- Hardy, J., and Singleton, A. (2009). Genomewide association studies and human disease. *New England Journal of Medicine* 360, 1759–1768.
- Hedrick, P. (2011). *Genetics of Populations* (4 ed). Jones & Bartlett Learning Publisher.
- Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genetics* 4, e1000008.
- Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology* 28, 500–544.
- Houle, D., Govindaraju, D.R., and Omholt, S.W. (2010). Phenomics: the next challenge. *Nature Reviews Genetics* 11, 855–866.

- Hunter, P.J. (2004). The IUPS Physiome Project: a framework for computational physiology. *Progress in Biophysics and Molecular Biology* 85, 551–569.
- Hunter, P.J., and Borg, T.K. (2003). Integration from proteins to organs: the Physiome Project. *Nature Reviews Molecular Cell Biology* 4, 237–243.
- Jacquard, A. (1983). Heritability: one word, three concepts. *Biometrics* 39, 465–477.
- Jaeger, J., Irons, D., and Monk, N. (2012). The Inheritance of Process: A Dynamical Systems Approach. *Journal of Experimental Zoology (Mol. Dev. Evol.)* 318, 591–612.
- Jansen, R.C., and Nap, J.P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics* 17, 388–391.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.L. (2000). The large-scale organization of metabolic networks. *Nature* 407, 651–654.
- Johnson, N.A.N., and Porter, A.H.A. (2000). Toward a new synthesis: population genetics and evolutionary developmental biology. *Genetica* 112-113, 45–58.
- Joyce, A.R., and Palsson, B.Ø. (2006). The model organism as a system: integrating “omics” data sets. *Nature Reviews Molecular Cell Biology* 7, 198–210.
- Kacser, H., and Burns, J.A. (1981). The molecular basis of dominance. *Genetics* 97, 639–666.
- Koerhuis, A. (1996). Non-normality of egg production distributions in poultry and the effects of outlier elimination and transformation on size and curvilinearity of heritability. *Livestock Production Science* 45, 69–85.
- Kohl, P., and Noble, D. (2009). Systems biology and the virtual physiological human. *Mol Syst Biol* 5, 292.
- Köhler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics* 82, 949–958.
- Lewontin, R.C. (1974). *The Genetic Basis of Evolutionary Change* (Columbia Biological Series). Columbia University Press.
- Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M.I., et al. (2010a). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology* 4, 92.
- Lloyd, C.M., Lawson, J.R., Hunter, P.J., and Nielsen, P.F. (2008). The CellML Model Repository. *Bioinformatics* 24, 2122–2123.
- Lotka, A.J. (1920). Undamped oscillations derived from the law of mass action. *Journal of The American Chemical Society* 42, 1595–1599.
- Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits* (1 ed.). Sinauer Associates.
- Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456, 18–21.

- Manolio, T.A., Collins, F.S., Cox, N.J., and Goldstein, D.B. (2009). Finding the missing heritability of complex diseases. *Nature* *461*, 747–753.
- McCarroll, S.A., and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics* *39*, S37–S42.
- Mestl, T., Plahte, E., and Omholt, S.W. (1995). A mathematical framework for describing and analysing gene regulatory networks. *Journal of Theoretical Biology* *176*, 291–300.
- Moore, J.H., and Williams, S.M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* *27*, 637–646.
- Moore, J.H.J. (2002). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity* *56*, 73–82.
- Nadeau, J.H., and Dudley, A.M. (2011). Systems Genetics. *Science* *331*, 1015–1016.
- Omholt, S.W. (2012). From sequence to consequence and back. *Progress in Biophysics and Molecular Biology* (in press).
- Omholt, S.W., Plahte, E., Oyehaug, L., and Xiang, K. (2000). Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* *155*, 969–980.
- Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., Markowitz, V.M., and Kyrpides, N.C. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* *40*, D571–D579.
- Peccoud, J., Velden, K.V., Podlich, D., Winkler, C., Arthur, L., and Cooper, M. (2004). The selective values of alleles in a molecular network model are context dependent. *Genetics* *166*, 1715–1725.
- Peck, A.L. (1943). Aristotle, *Generation of Animals*. In: Loeb Library Edition. Harvard University Press.
- Peng, B., and Amos, C.I. (2010). Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics* *11*, 442.
- 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. *Nature Reviews Genetics* *9*, 855–867.
- Plahte, E., Mestl, T., and Omholt, S.W. (1998). A methodological basis for description and analysis of systems with complex switch-like interactions. *Journal of Mathematical Biology* *36*, 321–348.
- Pumir, A., and Shraiman, B. (2011). Epistasis in a Model of Molecular Signal Transduction. *PLoS Computational Biology* *7*, e1001134.
- Rajasingh, H., Gjuvsland, A.B., Vage, D.I., and Omholt, S.W. (2008). When Parameters in Dynamic Models Become Phenotypes: A Case Study on Flesh Pigmentation in the Chinook Salmon (*Oncorhynchus tshawytscha*). *Genetics* *179*, 1113–1118.

Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. *Nat. Rev. Genet.* 7, 862–872.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.

Swami, M. (2009). Systems genetics: Networking complex traits. *Nat. Rev. Genet.* 10, 219–219.

Vik, J.O., Gjuvsland, A.B., Li, L., Tondel, K., Niederer, S., Smith, N.P., Hunter, P.J., and Omholt, S.W. (2011). Genotype-phenotype map characteristics of an in silico heart cell. *Frontiers in Genomic Physiology* 2, 106.

Vik, J.O., Gjuvsland, A.B., De Bono, B., and Omholt, S.W. (2012). From genotype to phenotype. In P. V. Convey, P. J. Hunter, M. Viceconti, D. Noble, & V. Diaz (Eds.), *Computational Biomedicine*. Oxford University Press Retrieved From [Http://Vph-Portal.eu/Vph-Textbook](http://Vph-Portal.eu/Vph-Textbook).

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. *Nature Reviews Genetics* 9, 255–266.

Volterra, V. (1926). *Variatione fluttuazioni del numero d'individui in specie conviventi*. *Mem. Acad. Lincei.* 2, 31–113.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* 6, 109–118.

Wang, Y., Gjuvsland, A.B., Vik, J.O., Smith, N.P., Hunter, P.J., and Omholt, S.W. (2012). Parameters in Dynamic Models of Complex Traits are Containers of Missing Heritability. *PLoS Computational Biology* 8, e1002459.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42, 565–569.

Zaitlen, N., and Kraft, P. (2012). Heritability in the genome-wide association era. *Human Genetics* 131, 1655–1664.

Zanzoni, A., Soler-López, M., and Aloy, P. (2009). A network medicine approach to human disease. *FEBS Letters* 583, 1759–1765.

Zeng, Z.B., Wang, T., and Zou, W. (2005). Modeling Quantitative Trait Loci and Interpretation of Models. *Genetics* 169, 1711–1725.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National*

Academy of Sciences of The United States of America *109*, 1193–1198.



# **Paper I**





# **On the Relationship between Heritability and Regulatory Architecture**

Yunpeng Wang<sup>1</sup>, Jon Olav Vik<sup>2</sup>, Stig W. Omholt<sup>1</sup>, Arne B. Gjuvsland<sup>1,\*</sup>

<sup>1</sup>*Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway,*

<sup>2</sup>*Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, N-1432 Ås, Norway,*

*\* Corresponding author. E-mail: arne.gjuvsland@umb.no*

## Abstract

Additive genetic variance ( $V_A$ ) and total genetic variance ( $V_G$ ) are core concepts in biomedical, evolutionary and production-biology genetics. What determines the large variation in reported  $V_A/V_G$  ratios from line-cross experiments, is not well understood. Here we report how the  $V_A/V_G$  ratio, and thus the ratio between narrow and broad sense heritability ( $h^2/H^2$ ), varies as a function of the regulatory architecture underlying genotype-to-phenotype (GP) maps. We studied five dynamic models (of the cAMP pathway, the glycolysis, the circadian rhythms, the cell cycle, and heart cell dynamics), assuming genetic variation in model parameters. Even when imposing purely linear genotype to parameter maps and no environmental variation we observe quite low  $V_A/V_G$  ratios and models with positive feedback and cyclic dynamics gave much lower  $V_A/V_G$  ratios than those without. Our results show that some regulatory architectures consistently maintain a transparent genotype-to-phenotype relationship, whereas other architectures generate more subtle patterns. Our approach can be used to elucidate these relationships across a whole range of biological systems in a systematic fashion.

## Author summary

The *broad-sense heritability* of a trait is the proportion of phenotypic variance attributable to genetic causes, while the *narrow-sense heritability* is the proportion attributable to additive gene effects. A better understanding of what underlies variation in the ratio of the two heritability measures, or the equivalent ratio of additive variance  $V_A$  to total genetic variance  $V_G$ , is important for production biology, biomedicine and evolution. We find that reported  $V_A/V_G$  values from line crosses vary greatly and ask if biological mechanisms underlying such differences can be elucidated by linking computational biology models with genetics. To this end we made use of models of the cAMP pathway, the glycolysis, circadian rhythms, the cell cycle and cardiocyte dynamics. We assumed additive gene action from genotypes to model parameters and studied the resulting GP maps and  $V_A/V_G$  ratios of system-level phenotypes. Our results show that some types of regulatory architectures consistently preserve a transparent genotype-to-phenotype relationship, whereas others generate more subtle patterns. Particularly, systems with positive feedback and cyclic dynamics resulted in more non-monotonicity in the GP map leading to lower  $V_A/V_G$

ratios. Our approach can be used to elucidate the  $V_A/V_G$  relationship across a whole range of biological systems in a systematic fashion.

## Introduction

The broad-sense heritability of a trait,  $H^2 = V_G/V_P$ , is the proportion of phenotypic variance attributable to genetic causes, while the narrow-sense heritability  $h^2 = V_A/V_P$ , is the proportion attributable to additive gene action. The non-additive genetic variance that distinguishes the two heritability measures has been subject to substantial controversy for more than 80 years (e.g., [1-6]). Through statistical arguments it was recently shown that for traits with many loci at extreme allele frequencies much of the genetic variance becomes additive with  $h^2/H^2$  or equivalently  $V_A/V_G$  typically  $>0.5$  [4]. In populations with intermediate allele frequencies, such as controlled line crosses, the picture becomes more nuanced [7]. Table 1 summarizes  $V_A/V_G$  ratios from a collection of studies on such populations, and it shows that the ratio can become very small in some cases. The explanation for the wide range of  $h^2/H^2$  ratios in line crosses must be sought in the genotype-phenotype (GP) map and the underlying biological systems rather than allele-frequencies.

It is important to understand the causal underpinnings of this observed variation in  $h^2/H^2$  ratios within and between biological systems for several reasons. In human quantitative genetics where twin studies are commonly used most heritability estimates refer to  $H^2$  [8] and in cases where  $h^2/H^2$  is low this can lead to unrealistic expectations about how much of the underlying causative variation may be located by linear QTL detection methods [6]. On the other hand, low narrow sense heritability for a given complex trait does not necessarily imply that the environment determines much of the variation. In evolutionary biology additive variance is the foremost currency for evolutionary adaptation and evolvability. Important questions in this context are for example (i) to which degree is there selection on the regulatory anatomies themselves to maintain high additive variance, (ii) are there organizational constraints in building adaptive systems such that in some cases a low  $h^2/H^2$  ratio must of necessity emerge while the proximal solution is still selected for? Moreover, in a production biology context in connection with genetic modifications of sexually reproducing organisms, one would like to ensure that the modifications would be passed over to future generations in a fully predictable way. Thus one would like to ensure that the modification becomes highly heritable in the narrow sense.

As a step towards a physiologically grounded understanding of the variation of the  $h^2/H^2$  relationship across biological systems or processes, we posed the question: Are there regulatory structures, or certain classes of phenotypes, more likely to generate low  $V_A/V_G$  ratios than others? Addressing this question requires the linking of genetic variation to computational biology in a population context (e.g., [9-18]), so-called causally-cohesive genotype-phenotype (cGP) modeling [13,17,18]. We applied this approach to five well-validated computational biology models describing, respectively, the glycolysis metabolic pathway in budding yeast [19], the cyclic adenosine monophosphate (cAMP) signaling pathway in budding yeast [20], the cell cycle regulation of budding yeast [21], the gene network underlying mammalian circadian rhythms [22], and the ion channels determining the action potential in mouse heart myocytes [23]. These models differ in their regulatory architecture; below, we show that they also differ in the range of  $V_A/V_G$  ratios that they can exhibit. In particular, positive feedback regulation and oscillatory behaviour seem to dispose for low  $V_A/V_G$  ratios. The results suggest that our approach can be used in a generic manner to probe how the  $h^2/H^2$  ratio varies as a function of regulatory anatomy.

## Methods

### Simulations of cGP models

The five cGP models were built and analyzed with the *cgptoolbox* (<http://github.com/jonvi/cgptoolbox>) an open-source Python package developed by the authors; further source code specific to the simulations in this paper is available on request. In the following we describe the three main parts of the workflow: (i) the mapping from genotypes to parameters, (ii) the mapping from parameters to phenotypes, i.e. solving the dynamic models and (iii) the setup of Monte-Carlo simulations combining the two mappings. For each model, we briefly describe its origins, the software used to solve it, which parameters were subject to genetic variation, what phenotypes were recorded, and criteria for omitting outlying datasets. Text S1 contains more detailed descriptions of all five models.

### Genotype to parameter mapping

For each model, the following procedure was repeated many times for different selections of parameters to be subjected to simulated genetic variation (see "Monte Carlo simulations" below for details). We started by sampling three polymorphic loci, each determining one or two parameter in the dynamic model. Tables of eligible loci

with corresponding parameters and their baseline values are listed in Table S1-5, corresponding to the cAMP, glycolysis, cell cycle, circadian and action potential models respectively. Heritable variation in a chosen parameter was generated for a single bi-allelic locus with allele indexes 0 and 1 in the following manner. First, two numbers  $r_1$  and  $r_2$  were sampled uniformly in the interval  $[0.7, 1.3]$ . The parameter value for a homozygote 00 was set to  $r_1 b$  where  $b$  is the baseline value, for a homozygote 11 the parameter value was  $r_2 b$ , while the heterozygous genotype 01 was assigned the average of the two homozygotes  $(r_1 + r_2)b/2$ .

### **cAMP model**

The model of the complete cAMP signaling pathway in *S. cerevisiae* [20] taking the external glucose level as input was downloaded as SBML code ([link](#)) and integrated using PySCeS [24]. Genetic variation was introduced on association/dissociation and phosphorylation/dephosphorylation rates of signal proteins (see Table S1). The initial steady state concentrations before adding external glucose, the peak values after adding glucose and the time taken to reach peak values of cellular proteins were recorded as phenotypes (Figure 1A for phenotype illustration and Table S6 for phenotype descriptions).

### **Glycolysis model**

The model published by Teusink et al. [19] describes glycolysis in *S. cerevisiae* through the kinetics of 13 glycolytic enzymes determining the fluxes of metabolite state variables. Genetic variation was introduced on maximal reaction rates for the enzymes (see Table S2). We downloaded the model from the BioModels database ([link](#)) in SBML L2 V1, and solved it with PySCeS [24] to find the stable steady state concentrations of metabolites, which used as phenotypes (see Figure 1B and Table S7). Datasets were discarded if one or more of the genotypes did not give a stable steady state, as can happen due to a saddle-node bifurcation [25].

### **Cell cycle model**

The consensus control mechanisms of the cell cycle in *S. cerevisiae* modeled by algebraic/differential equations that describe the continuous changes in state variables and discontinuities due cellular events [21] was obtained from the CellML repository ([link](#)). Genetic variation was introduced on the production and decay rates of various proteins (see Table S3). The published model contains reset rules (events) for both

parameters and state variables, but the CellML implementation only includes the parameter ( $k_{\text{mad}2}$ ,  $k_{\text{bub}2}$  and  $k_{\text{ite}1}$ ) rules. Reset rules for state variables [BUD], [SPN], and [ORI] as described in the model paper, were implemented by solving the model with rootfinding for the relevant variables. The model was numerically integrated using the CVODE solver [26] until convergence of cell division time, cell cycle events. The peak levels and time to peak levels for the cytosolic protein concentrations, together with the timing of cell division events were recorded as phenotypes (see Figure 1C for phenotype illustrations and Table S8 for phenotype descriptions).

### **Circadian model**

The mammalian circadian clock published by Leloup and Goldbeter [22,27] describes the dynamics mRNA and proteins of three genes in the cytosol and nucleus. Genetic variation was introduced on mRNA decay rates (see Table S4). The model was downloaded from CellML repository ([link](#)) and integrated using CVODE [26] until convergence to its limit cycle. As phenotypes we used the bottom levels, time to from bottom level to peak value of the concentrations of mRNAs, proteins and protein complexes. In addition, the time between two successive peak values are record as the period phenotype (see Figure 1D for phenotype illustrations and Table S9 for phenotype descriptions).

### **Action potential model**

The model of [28] describes the action potential and calcium transient of a mouse heart muscle cell. We obtained CellML code from the authors; numerical integration was done using CVODE [26]. Genetic variation was introduced on the maximal conductances of ion channels and pump affinities (see Table S5). Phenotypes were generated by simulated regular pacing as done in [17,18], with a stimulus potassium current of -15 V/s was lasting for 3 ms at the start of each stimulus interval. The model was simulated to convergence as described in [17] ; datasets that failed to converge were discarded. The initial level, peak level, amplitude, and time to 25, 50, 75 and 90% recovery of the action potential and calcium transient were recorded as the cell level phenotypes (see Figure 1E for phenotype illustrations and Table S10 for phenotype descriptions).

### **Monte Carlo simulations**

For each model we performed 1000 Monte Carlo simulations as follows.

We first sampled three polymorphic loci for introduction of genetic variation and sampled the genotype-to-parameter map as described above. Then all 27 possible three-locus genotypes were enumerated, mapped into 27 parameter sets and for each parameter set the dynamic model was solved and phenotypes (as described above and in Figure 1) were obtained. To avoid artifacts arising from numerical noise datasets with low variability were omitted from the genetic analysis. Absolute variability was measured as the difference between the maximum and minimum values of a phenotype in a dataset, and relative variability as the ratio of the absolute variation to the mean phenotype of the same dataset. The threshold values for each phenotype and the number of datasets exceeding the thresholds are listed in Tables S6-10, for the cAMP, glycolysis, cell cycle, circadian and action potential models, respectively.

## Statistical analysis

### Decomposition of genetic variance

A single Monte Carlo simulation results in genotype-to-phenotype maps comprised by 27 genotypic values (i.e. the phenotype values corresponding to the 27 genotypes) for a given phenotype. We used the NOIA framework [29] to compute the resulting genetic variance ( $V_G$ ) in a hypothetical  $F_2$  population and decompose it into additive ( $V_A$ ) and non-additive components. This was done with the function `linearGPmapanalysis` in the R package `noia` (<http://cran.r-project.org/web/packages/noia/>) version 0.94.1.

### Monotonicity of GP-maps

We build on the definitions of monotonicity and the indexing of alleles introduced in [30]. Given a simulated GP map with 27 genotypic values we measure the degree of order-breaking for a particular locus  $k$  by the allele substitution effects at that locus. For a fixed background genotype at all other loci (as indicated in eq. (14) in [30]) we compute the difference in genotypic value when substituting a 1-allele with a 2-allele (i.e. when going from 11 to 12 or from 12 to 22 at locus  $k$ ). We collect substitution effects across all 9 background genotypes compute  $N$ , the sum of all negative substitution effects and  $A$ , the sum of absolute values of all substitution effects. If the GP map is monotone for locus  $k$  then  $N=0$ . In Figure 5, only GP maps with  $N/A$  are counted as order-breaking.

## Results/Discussion

### System classification and phenotype dimensionality

The five cGP models studied in this work differ in many ways, both in their function and the underlying network structure. The glycolysis and cAMP models are both pathway models with an acyclic series of reactions transforming inputs to outputs. The glycolysis model [19] is more advanced than the metabolic models in earlier genetically oriented studies (e.g., [4,31,32]) in that it contains many different types of enzyme kinetics as well as negative feedback regulation of some enzyme activities through product inhibition. The cAMP model [20] contains a number of negative feedback loops, but overall it also has a clear pathway structure where the glucose signal is relayed from G-proteins to cAMP to the target kinase PKA. Both these two models have in common relatively simple dynamics with solutions converging to a stable steady state (Figure 1A and B).

In contrast, the three other models show cyclic behavior resulting from an interplay between positive and negative feedback loops (Figure 1 C-E). However, there are clear differences between these models too. The heart cell model [23] is an excitable system with feedback mechanisms such as calcium-induced calcium release and several voltage-dependent ion channels. In contrast to pacemaker cells, it relies on external pacing to initiate the action potential. The circadian rhythm model [22,27] is a gene expression network with intertwined positive and negative transcriptional feedback loops, giving a limit cycle oscillator with sustained oscillations even in continuous darkness. The cell cycle model [21] centers around a positive feedback loop between B-type cyclins in association with cyclin dependent kinase and inhibitors of the cyclin dependent kinase, which establishes a hysteresis loop causing the cell cycle to emerge from transitions between the two alternative stable steady states.

This crude classification of the five cGP models into pathway models and more complex regulatory systems is clearly reflected in the effective dimensionality of the phenotypes arising in our Monte Carlo simulations. We studied the phenotypic dimensionality for all five cGP models by principal components analysis for each Monte Carlo simulation (Figure S2). On average 95% of phenotypic variation of the glycolysis and cAMP models can be explained by the first 3 principal components, whereas, the cell cycle and heart cell models require the first 5 principal components and 7 is required for the circadian model. Since the genotype-to-parameter maps are analogous among the five models these differences in the effective dimensionality of



high-level phenotypes indicates that the mappings from parameters to phenotypes are simpler for the pathway models than the other three models. This, together with results on the effect of positive feedback in gene regulatory networks [33] suggested that the glycolysis and cAMP models might result in higher  $V_A/V_G$  ratios than the other three models.

### **The ratio of additive genetic variance to total genetic variance**

The  $V_A/V_G$  results confirmed the expectations. Furthermore, a number of distinct patterns emerged. The cAMP model shows the overall highest  $V_A/V_G$  ratios values (Figure 2A and Table S6), with mean and median values above 0.99 across all recorded phenotypes. The 0.05-quantile (i.e. only 5 percent of the Monte Carlo simulations show lower values than this)  $V_A/V_G$  values were above 0.97 for all phenotypes and no values lower than 0.6 were observed. In other words, the genotype-phenotype maps arising from this cGP model are very well approximated by an intra- and inter-locus additive model of gene action.

The glycolysis model too has mean and median  $V_A/V_G$  values close to 1 for all phenotypes (Figure 2B and Table S7), but compared to the cAMP model the numbers are clearly lower with *BPG* having the lowest recorded mean value of 0.9. The difference to the cAMP model is very clear when it comes to the lowest, with 0.05-quantiles below 0.7 for some of the phenotypes and values below 0.5 observed across all phenotypes. The distribution of  $V_A/V_G$  ratios for the cell cycle model (Figure S1 and Table S8) is quite similar to that of the glycolysis model with time to peak for *Sic1* having the lowest mean  $V_A/V_G$  values of 0.93 and 0.05-quantiles below 0.8 for some phenotypes.  $V_A/V_G$  values below 0.1 are observed for a few Monte Carlo simulations in some phenotypes.

For the cAMP, glycolysis and cell cycle models the distributions of  $V_A/V_G$  ratios were quite similar across all phenotypes and a large majority of the Monte Carlo simulations showed very high  $V_A/V_G$ . The circadian clock model differs from these both with large variation between phenotypes and with large proportions of the simulations ending up with low  $V_A/V_G$  (Figure 3A and Table S9). Four phenotypes stand out with  $V_A/V_G$  distributions that resemble a uniform distribution  $U(0,1)$ . These phenotypes are the time from bottom to peak for the phosphorylated and unphosphorylated proteins of *Per* and *Cry*, with median  $V_A/V_G$  ranging from 0.46 to 0.70 and 0.05-quantiles from 0.04 to 0.10. The remaining phenotypes give somewhat higher  $V_A/V_G$  values, but over half of the recorded phenotypes have 0.05-quantiles

below 0.6. For the action potential model all recorded phenotypes contain a large number of datasets showing low  $V_A/V_G$  ratios (Figure 3B and Table S10) with median values below 0.9 for the majority of phenotypes and with 0.05-quantiles in the range 0.18-0.35. The distributions are quite similar across action potential and calcium transient phenotypes, with the time to 90% repolarization for the action potential showing somewhat higher values than the others.

When comparing the distribution of  $V_A/V_G$  across the five cGP models a couple of remarks can be made. All five cGP models are capable of creating  $V_A/V_G$  ratios close to 1, and except for two phenotypes for the circadian model all median values of  $V_A/V_G$  are well above 0.5. This adds evidence in favor of the hypothesis [30] that inherent properties of biological systems could lead to considerable amounts of additive genetic variance even at intermediate allele frequencies. On the other hand, the incidence of low  $V_A/V_G$  ratios varied markedly among the five models, with the cAMP model giving virtually no non-additive genetic variance (Figure 2A), the glycolysis (Figure 2B) and cell cycle (Figure S1) models sometimes giving low  $V_A/V_G$  ratios, and the circadian and heart models doing so often (Figure 3). Encouragingly, the mean values of  $V_A/V_G$  obtained from our simulations cover the range of experimental values in Table 1. Because the genotype-parameter maps were purely additive, all non-additive genetic variance is a result of non-linearity in the ODE models. The expected effect of introducing non-additivity in the genotype-parameter maps would be a further decrease in the  $V_A/V_G$  ratios.

The association of positive feedback loops with lower  $V_A/V_G$  agrees with our earlier findings in gene regulatory networks [33]. A plausible explanation for the relatively high  $V_A/V_G$  ratios of the cell cycle model compared to the circadian and action potential models can be explained by a quote from Tyson and Novak's [34] discussion of why the cell-cycle is better understood as a hysteresis loop than as a limit cycle oscillator (LCO). "*Generally speaking, the period of an LCO is a complicated function of all the kinetic parameters in the differential equations. However, the period of the cell division cycle depends on only one kinetic property of the cell: its mass-doubling time.*" This could explain why the genotype-phenotype maps arising from the cell-cycle models is much more linear than the maps from the circadian model which is a LCO.

## **Systemic constraints on the genotype-phenotype map**

In populations with intermediate allele frequencies the  $V_A/V_G$  values are determined by the shape of the genotype-phenotype map, and differences between cGP models in

the distribution of  $V_A/V_G$  point to different constraints on the GP map. The concept of monotonicity of GP maps which we recently proposed [30] turns out to describe much of these constraints. Figure 4 depicts three extreme types of GP maps seen in our simulations. Nearly additive GP maps as shown in Figure 4A give  $V_A/V_G$  ratios very close to one. GP maps with strong magnitude epistasis as in Figure 4B result in intermediate  $V_A/V_G$  ratios while highly non-monotone GP maps (Figure 4C) showing strong overdominance and/or sign epistasis result in  $V_A/V_G$  ratios close to zero. Based on recent results from studies of gene regulatory networks [30] we expected that the three cGP models with complex regulation involving positive feedback should result in considerably more non-monotone or order-breaking GP maps than the two pathway models.

We measured the amount of order-breaking in all simulated GP maps (see Methods) and found a very clear pattern (Figure 5). While the datasets from the glycolysis and cAMP models contained only 1.1% and 1.3% GP maps with order-breaking for any locus, those from the cell cycle, circadian and action potential models contained 20.7%, 44.4% and 69.5%, respectively. We confirmed that monotone GP maps give higher  $V_A/V_G$  values than non-monotone GP maps for all five cGP models using the Mann-Whitney test (with p-values all below  $1e-10$ ).

The glycolysis model rarely shows order-breaking even for a single locus, but nevertheless can much lower  $V_A/V_G$  than the cAMP model. A plausible explanation is that the steady-state concentration of metabolites can increase strongly near a saddle-node bifurcation in the model [25]. Monte Carlo simulations with unstable steady states are discarded, but in those cases where one of the genotypes just comes close to the bifurcation point, we end up with genotype-phenotype maps as that seen in Figure 4B where one (or a small set of) genotype has much higher phenotypes than the others. Such GP maps, similar to the duplicate factor model in Hill et al [4], are known to give low  $V_A/V_G$  ratios despite being monotonic. Similar GP maps and  $V_A/V_G$  ratios close to zero were also found in the metabolic models studied by Keightley [32] in the case of null alleles at all loci.

### **Concluding remarks**

We have shown that the  $V_A/V_G$  ratio varies as a function of regulatory architecture and that distinct patterns emerge from linking computational biology models with genetic concepts. Our approach can be used to explain how properties of the GP map vary across and within biological systems. As it can be used with any computational

biology model, it has the potential to contribute substantially to a theoretical foundation capable of predicting when we are to expect low or high  $V_A/V_G$  or  $h^2/H^2$  ratios from the principles of regulatory biology. Causally cohesive genotype-phenotype modeling thus appears to qualify as a promising approach in the emerging field *systems genetics*, which aims to integrate causal models of biological networks with quantitative genetics [35-40].

## Acknowledgements

We are thankful to Katherine C. Chen for help on implementing and solving the cell cycle model.

## Reference

1. Fisher RA (1930) The genetical theory of natural selection. OxfordClarendon Press. 302p.
2. Wright S (1931) Evolution in Mendelian Populations. *Genetics* 16: 97–159.
3. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* 5: 618–625. doi:10.1038/nrg1407.
4. Hill WG, Goddard ME, Visscher PM (2008) Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet* 4: e1000008. doi:10.1371/journal.pgen.1000008.
5. Long N, Gianola D, Rosa GJM, Weigel KA (2011) Marker-assisted prediction of non-additive genetic values. *Genetica* 139: 843–854. doi:10.1007/s10709-011-9588-7.
6. Zuk O, Hechter E, Sunyaev SR, Lander ES (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.
7. Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* 10: 565–577. doi:10.1038/nrg2612.
8. Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sunderland: Sinauer Associates; 1 edition. 980p.
9. Omholt SW, Plahte E, Oyehaug L, Xiang K (2000) Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* 155: 969–980.
10. Peccoud J, Velden KV, Podlich D, Winkler C, Arthur L, et al. (2004) The selective values of alleles in a molecular network model are context dependent. *Genetics* 166: 1715–1725.

11. Welch SM, Dong Z, Roe JL, Das S (2005) Flowering time control: gene network modelling and the link to quantitative genetics : Modelling complex traits for plant improvement. *Aust J Agric Res* 56: 919–936.
12. Gjuvsland AB, Hayes BJ, Omholt SW, Carlborg O (2007) Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* 175: 411–420. doi:10.1534/genetics.106.058859.
13. Rajasingh H, Gjuvsland AB, Våge DI, Omholt SW (2008) When parameters in dynamic models become phenotypes: a case study on flesh pigmentation in the chinook salmon (*Oncorhynchus tshawytscha*). *Genetics* 179: 1113–1118. doi:10.1534/genetics.108.087064.
14. Gertz J, Gerke JP, Cohen BA (2010) Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions. *Theor Popul Biol* 77: 1–5. doi:10.1016/j.tpb.2009.10.002.
15. Salazar-Ciudad I, Jernvall J (2010) A computational model of teeth and the developmental origins of morphological variation. *Nature* 464: 583–586. doi:10.1038/nature08838.
16. Pumir A, Shraiman B (2011) Epistasis in a model of molecular signal transduction. *PLoS Comput Biol* 7: e1001134. doi:10.1371/journal.pcbi.1001134.
17. Vik JO, Gjuvsland AB, Li L, Tøndel K, Niederer S, et al. (2011) Genotype-Phenotype Map Characteristics of an In silico Heart Cell. *Front Physiol* 2: 106. doi:10.3389/fphys.2011.00106.
18. Wang Y, Gjuvsland AB, Vik JO, Smith NP, Hunter PJ, et al. (2012) Parameters in dynamic models of complex traits are containers of missing heritability. *PLoS Comput Biol* 8: e1002459. doi:10.1371/journal.pcbi.1002459.
19. Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, et al. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem* 267: 5313–5329.
20. Williamson T, Schwartz J-M, Kell DB, Stateva L (2009) Deterministic mathematical models of the cAMP pathway in *Saccharomyces cerevisiae*. *BMC Syst Biol* 3: 70. doi:10.1186/1752-0509-3-70.
21. Chen KC, Calzone L, Csikasz-Nagy A, Cross F, Novak B, et al. (2004) Integrative Analysis of Cell Cycle Control in Budding Yeast. *Mol Biol Cell* 15: 3841–3862. doi:10.1091/mbc.E03-11-0794.
22. Leloup J-C, Goldbeter A (2004) Modeling the mammalian circadian clock: sensitivity analysis and multiplicity of oscillatory mechanisms. *J Theor Biol* 230: 541–562. doi:10.1016/j.jtbi.2004.04.040.
23. Bondarenko VE, Szigeti GP, Bett GCL, Kim S-J, Rasmuson RL (2004) Computer model of action potential of mouse ventricular myocytes. *Am J Physiol Heart Circ Physiol* 287: H1378–H1403. doi:10.1152/ajpheart.00185.2003.
24. Olivier BG, Rohwer JM, Hofmeyr J-HS (2005) Modelling cellular systems

- with PySCeS. *Bioinformatics* 21: 560–561. doi:10.1093/bioinformatics/bti046.
25. Reijenga KA, van Megen YMGA, Kooi BW, Bakker BM, Snoep JL, et al. (2005) Yeast glycolytic oscillations that are not controlled by a single oscillator: a new definition of oscillator strength. *J Theor Biol* 232: 385–398. doi:10.1016/j.jtbi.2004.08.019.
  26. Cohen S, Hindmarsh C (1996) CVODE, a stiff/nonstiff ODE solver in C. *Computers in physics* 10: 138–143.
  27. Leloup J-C, Goldbeter A (2003) Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci U S A* 100: 7051–7056. doi:10.1073/pnas.1132112100.
  28. Li L, Niederer SA, Idigo W, Zhang YH, Swietach P, et al. (2010) A mathematical model of the murine ventricular myocyte: a data-driven biophysically based approach applied to mice overexpressing the canine NCX isoform. *Am J Physiol Heart Circ Physiol* 299: H1045–H1063. doi:10.1152/ajpheart.00219.2010.
  29. Alvarez-Castro JM, Carlborg O (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.
  30. Gjuvsland AB, Vik JO, Woolliams JA, Omholt SW (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.
  31. Kacser H, Burns JA (1981) The molecular basis of dominance. *Genetics* 97: 639–666.
  32. Keightley PD (1989) Models of quantitative variation of flux in metabolic pathways. *Genetics* 121: 869–876.
  33. Gjuvsland AB, Plahte E, Omholt SW (2007) Threshold-dominated regulation hides genetic variation in gene expression networks. *BMC Syst Biol* 1: 57. doi:10.1186/1752-0509-1-57.
  34. Tyson JJ, Novak B (2001) Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J Theor Biol* 210: 249–263. doi:10.1006/jtbi.2001.2293.
  35. Sieberts SK, Schadt EE (2007) Moving toward a system genetics view of disease. *Mamm Genome* 18: 389–401. doi:10.1007/s00335-007-9040-6.
  36. Rockman MV (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* 456: 738–744. doi:10.1038/nature07633.
  37. Benfey PN, Mitchell-Olds T (2008) From genotype to phenotype: systems biology meets natural variation. *Science* 320: 495–497. doi:10.1126/science.1153716.
  38. Flint J, Mackay TFC (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* 19: 723–733. doi:10.1101/gr.086660.108.

39. Nadeau JH, Dudley AM (2011) Genetics. Systems genetics. *Science* 331: 1015–1016. doi:10.1126/science.1203869.
40. Moore A (2012) Towards the new evolutionary synthesis: Gene regulatory networks as information integrators. *Bioessays* 34: 87–87. doi:10.1002/bies.201290000.
41. Le Rouzic A, Alvarez-Castro JM, Carlborg O (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.
42. Prows DR, Hafertepen AP, Gibbons WJ, Winterberg AV, Nick TG (2007) A genetic mouse model to investigate hyperoxic acute lung injury survival. *Physiol Genomics* 30: 262–270. doi:10.1152/physiolgenomics.00232.2006.
43. Jordan KW, Carbone MA, Yamamoto A, Morgan TJ, Mackay TFC (2007) Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome Biol* 8: R172. doi:10.1186/gb-2007-8-8-r172.
44. Wang P, Lyman RF, Shabalina SA, Mackay TFC, Anholt RRH (2007) Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in *Drosophila*. *Genetics* 177: 1655–1665. doi:10.1534/genetics.107.079731.
45. Zhang JF, Waddell C, Sengupta-Gopalan C, Potenza C, Cantrell RG (2007) Diallel analysis of root-knot nematode resistance based on galling index in upland cotton. *Plant Breeding* 126: 164–168. doi:10.1111/j.1439-0523.2007.01303.x.
46. Hao JJ, Yu SX, Dong ZD, Fan SL, Ma QX, et al. (2008) Quantitative inheritance of leaf morphological traits in upland cotton. *J Agric Sci* 146: 561–569.
47. Cuevas HE, Song H, Staub JE, Simon PW (2009) Inheritance of beta-carotene-associated flesh color in cucumber (*Cucumis sativus* L.) fruit. *Euphytica* 171: 301–311. doi:10.1007/s10681-009-0017-2.
48. Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM (2009) Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS ONE* 4: e7433. doi:10.1371/journal.pone.0007433.
49. Lyimo HJF, Pratt RC, Mnyuku RSOW (2011) Heritability and gene effect estimates for components of partial resistance to grey leaf spot of maize by generation mean analysis\*. *Plant Breeding* 130: 633–639. doi:10.1111/j.1439-0523.2011.01890.x.
50. Kearsley MJ, Pooni HS, Syed NH (2003) Genetics of quantitative traits in *Arabidopsis thaliana*. *Heredity* 91: 456–464. doi:10.1038/sj.hdy.6800306.
51. Chakravarthi DVN, Rao YV, Rao MVS, Manga V (2010) Genetic analysis of in vitro callus and production of multiple shoots in eggplant. *Plant Cell Tiss Organ Cult* 102: 87–97. doi:10.1007/s11240-010-9709-5.





## Figure legends

**Figure 1. Phenotypes derived from the cGP models.** Graphical illustration of the phenotypes recorded for the five cGP models studied. Time courses (state variable on y-axis, time on x-axis) for the baseline parameter set are displayed for all five models. **A.** In the absence of external glucose all state variables (concentration of cAMP is shown) in the cAMP model [20] converge to a stable steady state (blue circle on y-axis). After addition of external glucose (5mM added at time 50) we see a rapid change followed by a slow return to the original steady state. In addition to the original steady state, the extremal concentration (top blue circle) as well as the time to reach the extremum (blue line) was recorded as phenotypes. **B.** Metabolite concentrations (internal glucose (GLCi), glucose-6-phosphate (G6P) and fructose-6-phosphate (F6P) are shown) in the glycolysis model [19] all converge to a stable steady state, indicated by open circles. The steady state concentrations for 13 metabolites were recorded as phenotypes from this model. **C.** For the cell cycle model [21] we recorded the peak level and the time from bottom to peak as for the circadian model (Figure 1D), and in addition we recorded cell cycle events such as bud emergence at the time when [BUD]=1 indicated by the black arrow. **D.** mRNA and protein concentrations (mRNA for *Bmal1* (MB), *Cry* (MC) and *Per* (MP) are shown) in the circadian model [22] converge to a limit cycle. In addition to the period of oscillation (long blue line) for each of the 16 variables the peak level (open blue circle) as well as the time from bottom to peak (short blue line) were recorded as phenotypes. **E.** We used the base level, peak level, amplitude, time to peak, and time to 25%, 50%, 75% and 90% recovery of the action potential and calcium transient as cell level phenotypes of the action potential model. An action potential is shown in the figure.

**Figure 2. The empirical cumulative distribution function of  $V_A/V_G$  ratios for phenotypes of the cAMP (A) and the glycolysis (B) models.** **A.** The empirical cumulative distribution functions (y axis) of  $V_A/V_G$  ratios (x axis) for all phenotypes studied in the cAMP model: The initial steady state concentrations before adding external glucose of the cyclic adenosine monophosphate (cAMP), the G-protein Ras2a (Ras2a), the guanine-nucleotide-exchange factor (Cdc25), the protein kinase A (PKAi). The peak values after adding glucose of these proteins (cAMPv, Ras2av, Cdc25v and PKAiv), the Kelch repeat homologue protein (Krhv), the G-protein Gpa2a (Gpa2av), and the phosphodiesterase (Pde1v). The time taken to reach the

peak values (cAMPt, Ras2at, Cdc25t, PKAit, Krht, Gpa2at, Ped1t). See Table S6 for further phenotype descriptions and numerical summaries of the distribution of  $V_A/V_G$  ratios. **B.** The empirical cumulative distribution function (y axis) of  $V_A/V_G$  ratios (x axis) for the steady state concentrations of 13 metabolites in the glycolysis model: acetaldehyde (ACE), 1,3-bisphosphoglycerate (BPG), fructose-1,6-bisphosphate (F16P), fructose 6-phosphate (F6P), glucose 6-phosphate (G6P), glucose in cell (GLCi), nicotinamide adenine dinucleotide (NADH), phosphates in adenine nucleotide (P), 2-phosphoglycerate (P2G), 3-phosphoglycerate (P3G), phosphoenolpyruvate (PEP), pyruvate (PYP), and trio-phosphate (TRIO). See Table S7 for further phenotype descriptions and numerical summaries of the distribution of  $V_A/V_G$  ratios.

**Figure 3. The empirical cumulative distribution function of  $V_A/V_G$  ratios for phenotypes of the circadian model (A) and the action potential model (B).** The empirical cumulative distribution functions (y axis) of  $V_A/V_G$  ratios (x axis) for phenotypes studied in the circadian model and the heart cell model. **A.** The upper-left panel (Bmal1) shows phenotypes related to *bmal1* gene, including the mRNA (MB), the unphosphorylated/phosphorylated protein in cytosol (BC/BCP) and nucleus (BN/BNP). The upper-right panel (Complex) is for protein complexes PCC, PCN, PCCP and PCNP. The bottom-right panel (Per) is for *per* gene, including the mRNA (MP), the unphosphorylated protein (PC) and the phosphorylated protein (PCP). The bottom concentration (solid line) and the time take to peak (dashed line) of each species are recorded phenotypes. The bottom-left panel (Cry) is related to *cry* gene, including the mRNA (MC), the unphosphorylated protein (CC) and phosphorylated protein (CCP). The period of circadian rhythm (Period, dotted line) is also shown. See Table S9 for further phenotype descriptions and numerical summaries of the distribution of  $V_A/V_G$  ratios. **B.** The empirical cumulative distribution functions (y axis) of  $V_A/V_G$  ratios (x axis) for phenotypes studied in the action potential model: time to 25%, 50%, 75% and 90% of initial values, the amplitude, initial values (Base), peak values, time to reach peak values of action potential (left panel) and calcium transient (right panel) are shown. See Table S10 for further phenotype descriptions and numerical summaries of the distribution of  $V_A/V_G$  ratios.

**Figure 4. Three distinct types of genotype-phenotype maps.** Examples of three distinct types of genotype-phenotype maps seen in our simulations. **A.** A nearly additive map exemplified by the GP map of the time to peak concentration of Cdc25 ( $V_A/V_G = 0.99$ ) in the cAMP model; **B.** A fully monotone but non-additive map exemplified by the GP map of the concentration of P2G protein ( $V_A/V_G = 0.41$ ) of the Glycolysis model; and, **C.** A strongly non-monotonic map is found the time to peak concentration of the PC protein ( $V_A/V_G = 0.03$ ) from the circadian model. For each panel the phenotypic value is shown on each y-axis while the x-axes, line colours and plot columns indicate the genotype at the three loci.

**Figure 5. The number of loci for which the GP-map shows order-breaking.** The number of Monte Carlo simulations where the GP-map for a given phenotype is clearly order-breaking (see Methods) is shown for the cAMP model (**A**), the glycolysis model (**B**), the cell cycle model (**C**), the circadian model (**D**) and the action potential model (**E**). Only phenotypes with at least one out of the 1000 simulated GP map showing order-breaking for at least one locus, are shown.

## Tables

**Table 1. Summary of  $V_A/V_G$  Ratios of traits in line-crossing experiment reported from other groups.**

Species	Type of traits	Number of traits	Min $V_A/V_G$	Max $V_A/V_G$	Mean $V_A/V_G$
Chicken	Bodyweight <sup>a</sup> [41]	17	0.03	0.71	0.34
Mouse	Hyperoxic Survival [42]	1	-	-	0.46
<i>Drosophila melanogaster</i>	Locomotor Behavior[43]	1	-	-	0.31
	Olfactory Behavior[44]	1	-	-	0.64
Root-knot nematode	Resistance in the Auburn and Nem-X resistance sources [45]	1	-	-	0.79
Cotton	Morphological traits <sup>b</sup> [46]	6	0.0	1	0.48
Melon	Beta-carotene-associated mesocarp color <sup>c</sup> and fruit maturity[47]	2	0.55	0.58	0.57
Maize	Morphological traits <sup>d</sup> [48]	17	0.13	1.1	0.61
	Partial resistance to grey leaf spot <sup>e</sup> [49]	5	0.51	0.92	0.76
<i>Arabidopsis</i>	Morphological traits <sup>f</sup> [50]	22	0.58	1.05	0.76

<i>thaliana</i>					
Eggplant	Callus related traits <sup>g</sup> [51]	4	0.42	0.92	0.73

- a. Averaged over measures of bodyweight and growth rate.
- b. Averaged over both lines (I and II) and leaf morphological traits.
- c. Averaged over measures based on generation mean method, variance component method on individuals and on families.
- d. Averaged over both traits and inbred and hybrid lines, and For each trait the average narrow sense heritability was used.
- e. Average of three crosses and five resistance phenotypes.
- f. The ratio of each trait was calculated by  $h^2_n/h^2_b$  of the triple test crosses and the average over all traits is shown.
- g. Averaged over four traits.

## Supporting Information

### Text S1 More detailed descriptions of the five cGP models.

**Figure S1. The empirical cumulative distribution function of  $V_A/V_G$  ratios for phenotypes of the cell cycle model.** The empirical cumulative distribution functions (y axis) of  $V_A/V_G$  ratios (x axis) for all phenotypes studied in the cell cycle model. The phenotypes are divided into 3 groups. Cell events refer to the discrete events defined in the model paper and include timing of budding (Bud), timing of DNA replication (Rep) and timing of alignment of chromosomes on the metaphase plates (Spn). Peak concentration include the concentration of the phosphorylated anaphase-promoting complex (APCP), the G1-stabilizing protein Cdc6, the B-type Cyclin protein 2 (Clb2), the S-phase promoting B-type Cyclin (Clb5), the starter kinase (Cln2) and the G1 phase stabilizing protein (Sci1). The time to peak phenotypes include the time to reach peak concentrations of APCP, Cdc6, Clb2, Clb5, Cln2 and Sci1. See Table S8 for further phenotype descriptions and numerical summaries of the distribution of  $V_A/V_G$  ratios.

**Figure S2. Dimensionality of phenotypic variation.** The percentage of total phenotypic variation (y axis) versus the number of principal components (x axis) across all five cGP models (colour coded). For each Monte Carlo data set the  $27 \times M$  matrices containing the full genotype-phenotype map for all  $M$  recorded phenotypes

was standardized to unit variance before principal components analysis. Each boxplot summarizes explained variance for 1000 or close to 1000 simulations.

**Table S1. Polymorphic model elements of the cAMP model.** A list of cAMP model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

**Table S2. Polymorphic model elements of the glycolysis model.** A list of glycolysis model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

**Table S3. Polymorphic model elements of the cell cycle model.** A list of cell cycle model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

**Table S4. Polymorphic model elements of the circadian model.** A list of circadian model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

**Table S5. Polymorphic model elements of the action potential model.** A list of action potential model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

**Table S6. Summary of phenotype descriptions, variability thresholds and distribution of  $V_A/V_G$  ratios for the cAMP model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

**Table S7. Summary of phenotype descriptions, variability thresholds and distribution of  $V_A/V_G$  ratios for the glycolysis model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes

and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

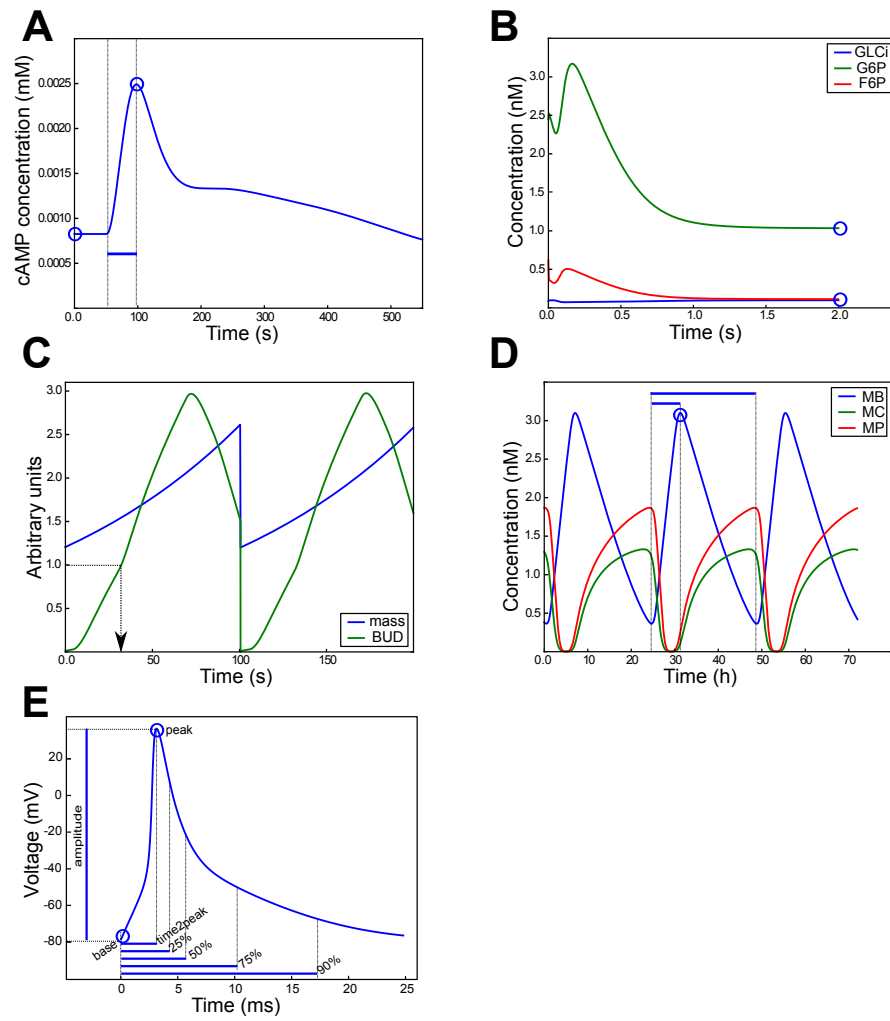
**Table S8. Summary of phenotype descriptions, variability thresholds and distribution of  $V_A/V_G$  ratios for the cell cycle model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

**Table S9. Summary of phenotype descriptions, variability thresholds and distribution of  $V_A/V_G$  ratios for the circadian model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold. Abbreviations: phosphorylated – phos., cytosolic – cyt., nuclear – nuc., bottom concentration – b.c., peak concentration – p.c.

**Table S10. Summary of phenotype descriptions, variability thresholds and distribution of  $V_A/V_G$  ratios for the action potential model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.



**Figure 1**



**Figure 2**

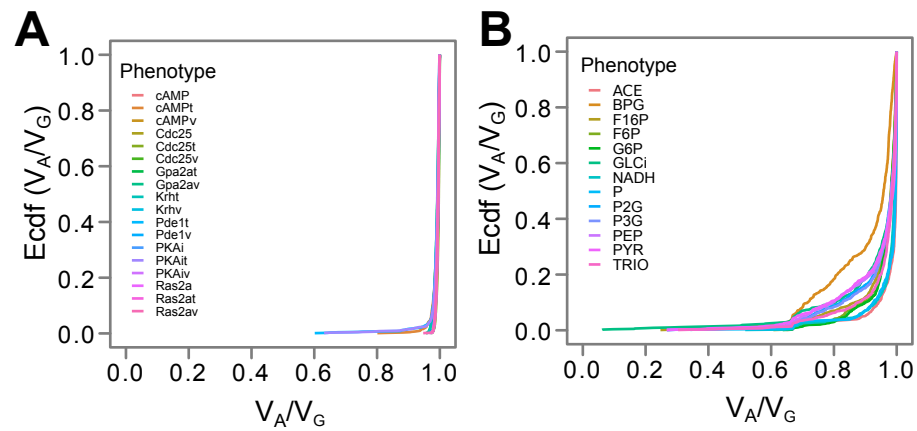




Figure 3

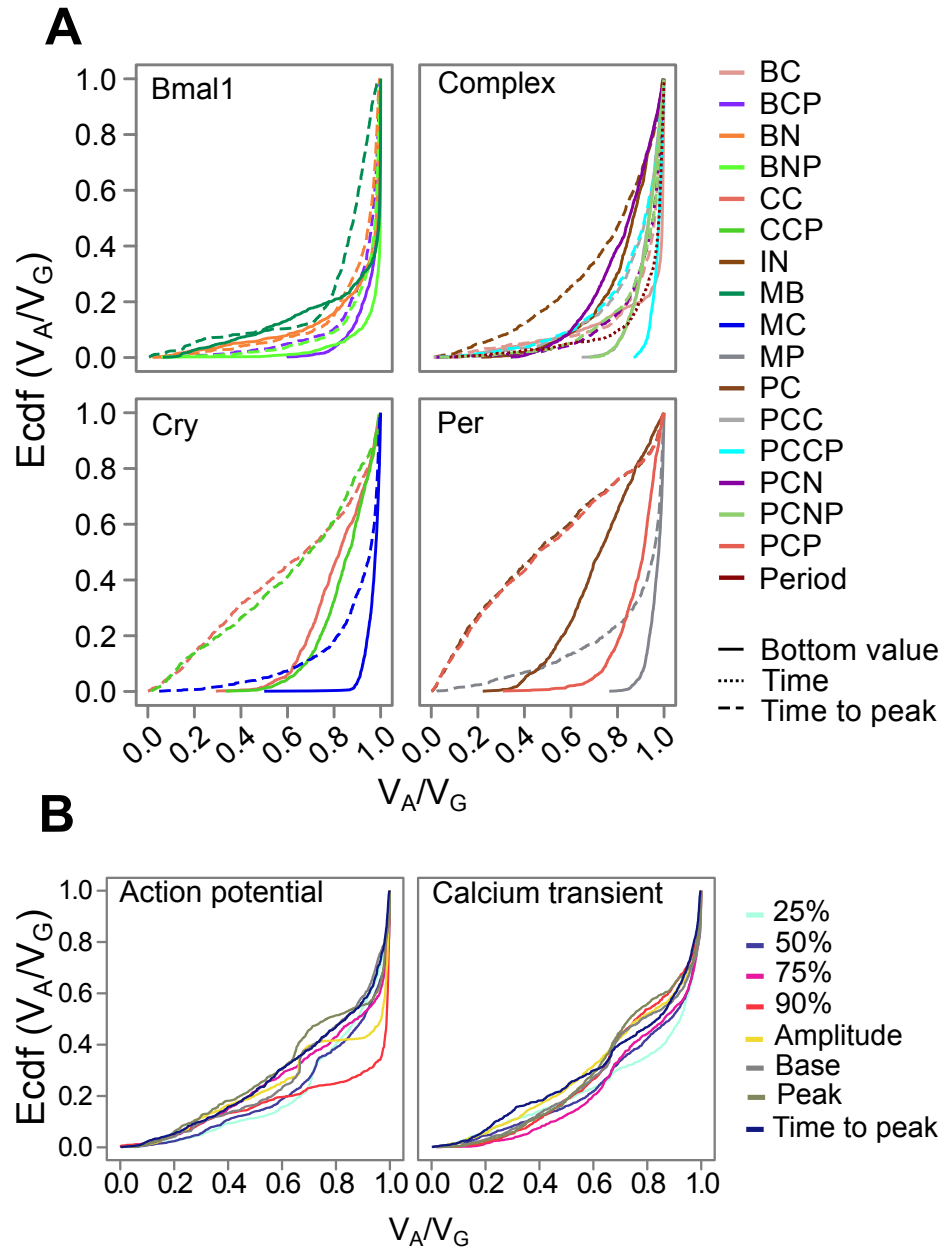
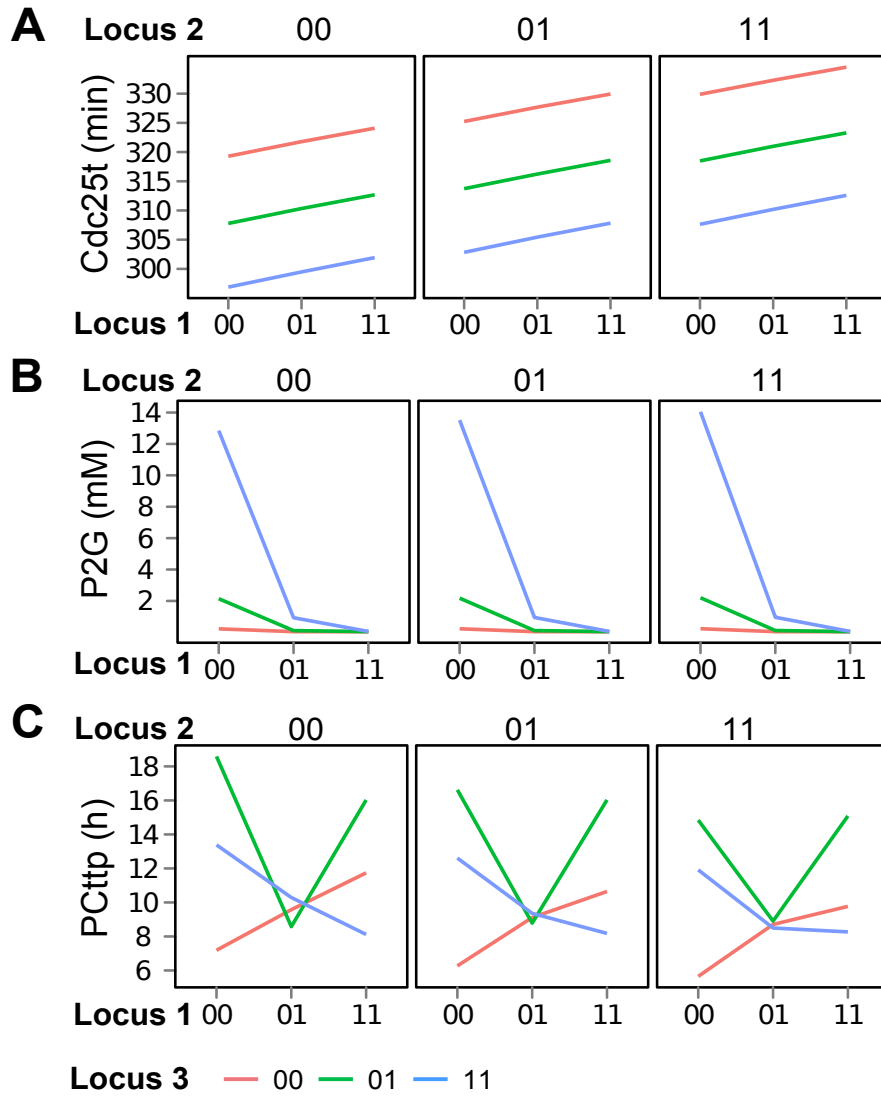


Figure 4



**Figure 5**

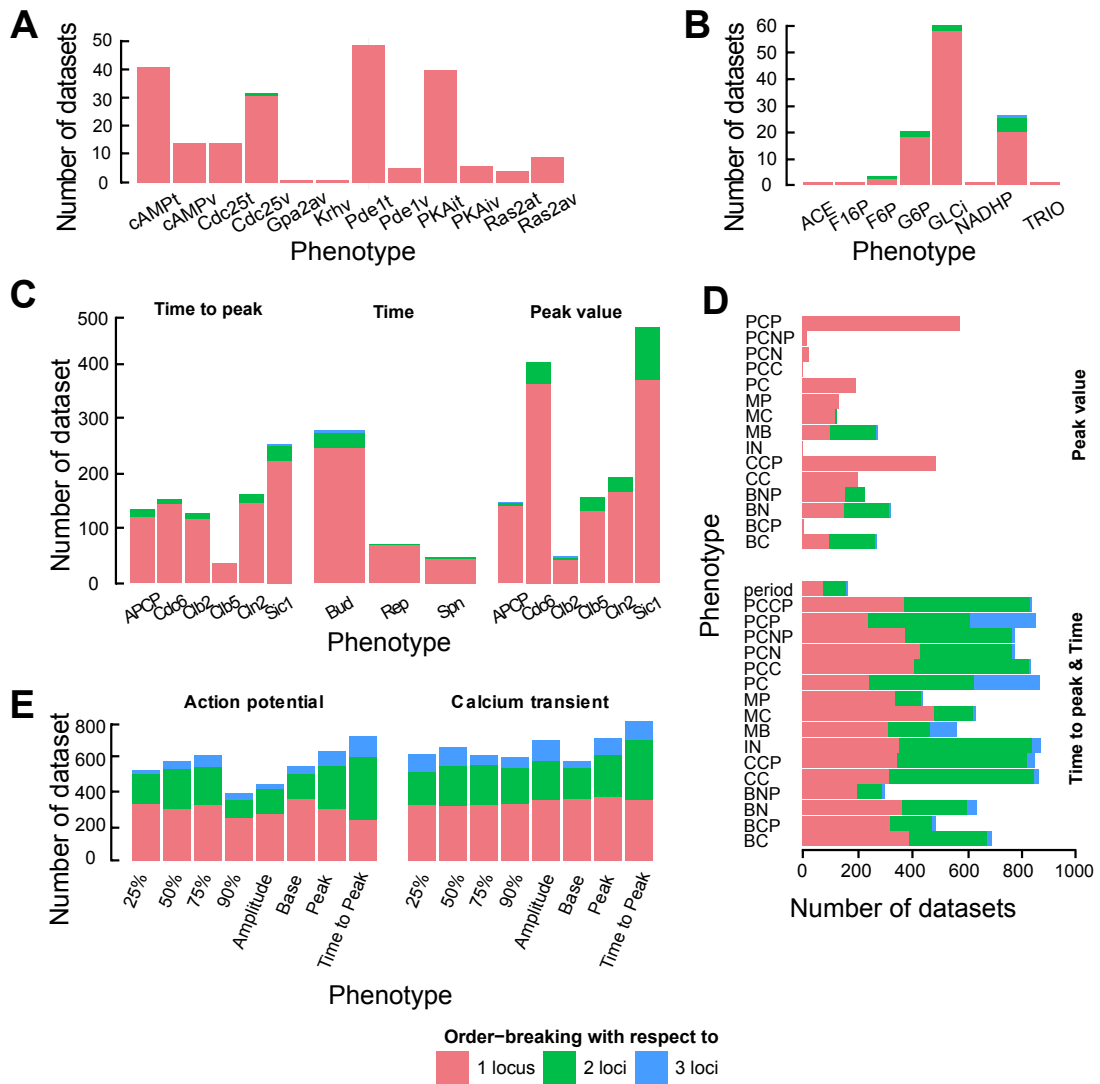


Figure S1

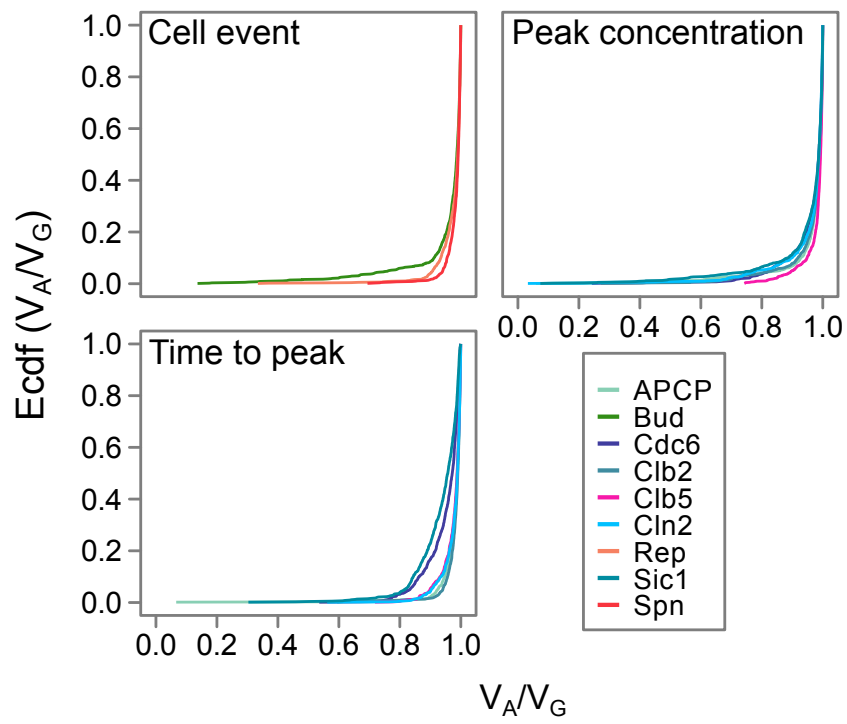
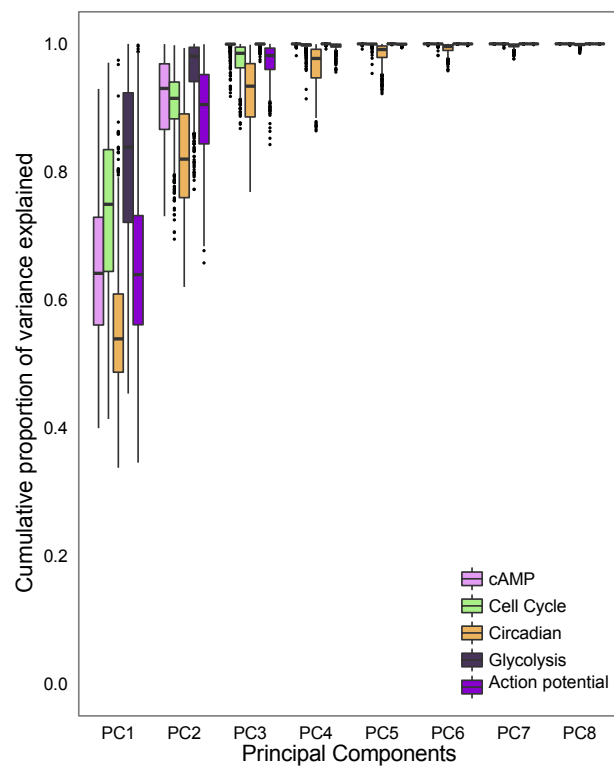


Figure S2



## On the relationship between heritability and regulatory architecture

Yunpeng Wang, Jon Olav Vik, Stig W. Omholt, Arne B. Gjuvsland\*

**Table S1. Polymorphic model elements of the cAMP model.** A list of cAMP model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

<b>Model element</b>	<b>Parameters</b>	<b>Baseline values</b>
1. G-protein coupled glucose receptor with glucose	Association rate	0.003 $l^2/mM*s$
	Dissociation rate	0.143 1/s
2. Kelch Repeat Homologue protein with the G-protein Gpa2	Association rate	391089.57 $l^2/mM*s$
	Dissociation rate	6.122 $l^2/mM*s$
3. G-protein Gpa2	Activation rate	57682.62 $l^2/mM*s$
	Deactivation rate	12989.422 $l^2/mM*s$
4. G-protein Ras2	Activation rate	0.74 1/s
	Deactivation rate	0.042 1/s
5. Guanine-nucleotide-exchanger factor	Phosphorylation rate	0.18 1/s
	Dephosphorylation rate	2.52 1/s
6. Phosphodiesterase	Phosphorylation rate	6.82 1/s
	Dephosphorylation rate	2.4 1/s

**Table S2. Polymorphic model elements of the glycolysis model.** A list of glycolysis model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

<b>Model element</b>	<b>Parameters</b>	<b>Baseline values</b>
1. Hexokinase	$V_{max}$	226.452 mM/min
2. Glucose-6-phosphate isomerase	$V_{max}$	339.677 mM/min
3. Phosphofructokinase	$V_{max}$	182.903 mM/min
4. Aldolase	$V_{max}$	322.258 mM/min
5. Glyceraldehyde 3-phosphate dehydrogenase	$V_{max}$	1184.52 mM/min
6. Phosphoglycerate mutase	$V_{max}$	2525.81 mM/min
7. Phosphoglycerate kinase	$V_{max}$	1306.45 mM/min
8. Enolase	$V_{max}$	365.806 mM/min
9. Pyruvate decarboxylase	$V_{max}$	174.194 mM/min
10. Glucose transport	$V_{max}$	97.264 mM/min
11. Alcohol dehydrogenase	$V_{max}$	810 mM/min
12. Glyceral 3-phosphate dehydrogenase	$V_{max}$	70.15 mM/min
13. Pyruvate kinase	$V_{max}$	1088.71 mM/min

**Table S3. Polymorphic model elements of the cell cycle model.** A list of cell cycle model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

<b>Model species</b>	<b>Parameters</b>	<b>Baseline values</b>
1. <i>Cln3</i>	$C_0$	0.4 dimensionless
2. <i>Bck2</i>	$B_0$	0.054 dimensionless
3. <i>Cln2</i>	Production rate	0.15 min <sup>-1</sup>
	Decay rate	0.12 min <sup>-1</sup>
4. <i>Clb5</i>	Production rate	0.005 min <sup>-1</sup>
	Decay rate	0.01 min <sup>-1</sup>
5. <i>Clb2</i>	Production rate	0.04 min <sup>-1</sup>
	Decay rate	0.003 min <sup>-1</sup>
6. <i>Cdc15</i>	Production rate	1 min <sup>-1</sup>
	Decay rate	0.5 min <sup>-1</sup>
7. <i>Cdc14</i>	Production rate	0.2 min <sup>-1</sup>
	Decay rate	0.1 min <sup>-1</sup>
8. <i>Pds1</i>	Production rate	0.03 min <sup>-1</sup>
	Decay rate	0.01 min <sup>-1</sup>

**Table S4. Polymorphic model elements of the circadian model.** A list of circadian model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

<b>Model element</b>	<b>Parameters</b>	<b>Baseline values</b>
1. <i>per</i>	mRNA decay rate	1.1 nM/h
2. <i>bmal1</i>	mRNA decay rate	0.2 nM/h
3. <i>cry</i>	mRNA decay rate	1.0 nM/h

**Table S5. Polymorphic model elements of the action potential model.** A list of action potential model elements used as polymorphic loci together with the names and baseline values of parameters used to manifest genetic variation.

<b>Model element</b>	<b>Parameters</b>	<b>Standard values</b>
1. SERCA	affinity to $\text{Ca}^{2+}$	$4.12 \times 10^{-1}$ uM
2. L-type $\text{Ca}^{2+}$ channel	permeability	$2.5 \text{ ms}^{-1}$
3. Calsequestrin	affinity to $\text{Ca}^{2+}$	$6.30 \times 10^2$ uM
4. $\text{Na}^+$ channel	maximum conductance	16 ms/uF
5. ultrarapidly activating delayed rectifier $\text{K}^+$ channel	The maximum conductance	0.25 ms/uF
6. rapidly recovering transient outward $\text{K}^+$ channel	maximal conductance	0.53 ms/uF
7. time-dependent $\text{K}^+$ channel	maximal conductance	0.35 ms/uF
8. $\text{Na}^+/\text{K}^+$ pump	affinity of the to intracellular $\text{Na}^+$	$1.66 \times 10^4$ uM



**Table S6. Summary of phenotypic values, variability thresholds and distribution of  $V_A/V_G$  ratios for the cAMP model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

Phenotype	Description	Units	Variability threshold		# of valid datasets	Quantiles and mean values of $V_A/V_G$						
			rel.	abs.		Q <sub>0.05</sub>	Q <sub>0.1</sub>	Q <sub>0.2</sub>	Q <sub>0.3</sub>	Q <sub>0.5</sub>	Q <sub>0.8</sub>	mean
<b>cAMP</b>	cAMP steady state concentration ( <i>SSC</i> ) before adding glucose	mM	0.01	1e-8	866	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>cAMPt</b>	Time to peak concentration ( <i>PC</i> ) of cAMP	min	0.01	1	879	0.97	0.98	0.98	0.99	0.99	1	0.99
<b>cAMPv</b>	<i>PC</i> of cAMP after adding glucose	mM	0.01	1e-8	983	0.98	0.99	0.99	0.99	1	1	0.99
<b>Cdc25</b>	<i>SSC</i> of Cdc25 before adding glucose	mM	0.01	1e-8	838	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>Cdc25t</b>	Time to <i>PC</i> of Cdc25	min	0.01	1	986	0.99	0.99	0.99	0.99	1	1	1
<b>Cdc25v</b>	<i>PC</i> of Cdc25 after adding glucose	mM	0.01	1e-8	951	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>Gpa2at</b>	Time to <i>PC</i> of G- protein Gpa2a	min	0.01	1	782	0.98	0.98	0.98	0.99	0.99	1	0.99
<b>Gpa2av</b>	<i>PC</i> of Gpa2a after adding glucose	mM	0.01	1e-8	890	0.98	0.99	0.99	0.99	1	1	0.99
<b>Krht</b>	Time to <i>PC</i> of Krh	min	0.01	1	785	0.98	0.98	0.98	0.99	0.99	1	0.99
<b>Krhv</b>	<i>PC</i> of Krh after adding glucose	mM	0.01	1e-8	803	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>Pde1t</b>	Time to <i>PC</i> of Pde1	min	0.01	1	961	0.97	0.98	0.99	0.99	0.99	1	0.99
<b>Pde1v</b>	<i>PC</i> of Pde1 after adding glucose	mM	0.01	1e-8	724	0.98	0.98	0.99	0.99	0.99	1	0.99
<b>PKAi</b>	<i>SSC</i> of PKA before adding glucose	mM	0.01	1e-8	622	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>PKAit</b>	Time to <i>PC</i> of Protein kinase A	min	0.01	1	964	0.97	0.98	0.99	0.99	0.99	1	0.99
<b>PKAiv</b>	<i>PC</i> of PKA after adding glucose	mM	0.01	1e-8	998	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>Ras2a</b>	<i>SSC</i> of Ras2a before adding glucose	mM	0.01	1e-8	846	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>Ras2at</b>	Time to <i>PC</i> of Ras2a	min	0.01	1	937	0.98	0.99	0.99	0.99	0.99	1	0.99
<b>Ras2av</b>	<i>PC</i> of Ras2a after adding glucose	mM	0.01	1e-8	865	0.98	0.99	0.99	0.99	1	1	0.99

**Table S7. Summary of phenotypic values, variability thresholds and distribution of  $V_A/V_G$  ratios for the glycolysis model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

Phenotype	Description	Units	Variability threshold		# of valid datasets	Quantiles and mean values of $V_A/V_G$						
			rel.	abs.		Q <sub>0.05</sub>	Q <sub>0.1</sub>	Q <sub>0.2</sub>	Q <sub>0.3</sub>	Q <sub>0.5</sub>	Q <sub>0.8</sub>	mean
<b>ACE</b>	Steady state concentration (SSC) of acetaldehyde		0.01	1e-4	728	0.90	0.94	0.98	0.99	1	1	0.98
<b>BPG</b>	SSC of bisphosphoglycerate		0.01	1e-4	478	0.67	0.72	0.81	0.90	0.96	0.98	0.90
<b>F16P</b>	SSC of fructose-1,6-bisphosphate		0.01	1e-4	934	0.75	0.87	0.94	0.96	0.98	1	0.95
<b>F6P</b>	SSC of fructose 6-phosphate		0.01	1e-4	908	0.83	0.89	0.95	0.96	0.98	1	0.96
<b>G6P</b>	SSC of glucose 6-phosphate		0.01	1e-4	922	0.84	0.90	0.95	0.96	0.98	1	0.96
<b>GLCi</b>	SSC of internal glucose in cell		0.01	1e-4	794	0.68	0.81	0.91	0.95	0.99	1	0.93
<b>NADH</b>	SSC of nicotinamide adenine dinucleotide		0.01	1e-4	710	0.88	0.93	0.97	0.98	0.99	1	0.97
<b>P</b>	SSC of phosphates in adenine nucleotides		0.01	1e-4	665	0.88	0.93	0.97	0.99	1	1	0.97
<b>P2G</b>	SSC of 2- phosphoglycerate		0.01	1e-4	896	0.73	0.82	0.93	0.96	0.98	1	0.94
<b>P3G</b>	SSC of 3-phosphoglycerate		0.01	1e-4	919	0.73	0.83	0.93	0.96	0.98	1	0.95
<b>PEP</b>	SSC of phosphoenolpyruvate		0.01	1e-4	886	0.71	0.80	0.92	0.96	0.98	1	0.94
<b>PYR</b>	SSC of pyruvate		0.01	1e-4	788	0.68	0.79	0.91	0.95	0.99	1	0.94
<b>TRIO</b>	SSC of trio-phosphate		0.01	1e-4	851	0.77	0.88	0.95	0.97	0.99	1	0.96

**Table S8. Summary of phenotypic values, variability thresholds and distribution of  $V_A/V_G$  ratios for the cell cycle model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

Phenotype	Description	Units	Variability threshold		# of valid datasets	Quantiles and mean values of $V_A/V_G$						
			rel.	abs.		Q <sub>0.05</sub>	Q <sub>0.1</sub>	Q <sub>0.2</sub>	Q <sub>0.3</sub>	Q <sub>0.5</sub>	Q <sub>0.8</sub>	mean
<b>APCP</b>	Peak concentration ( <i>PC</i> ) of phosphorylated anaphase-promoting complex	au	0.01	1e-4	948	0.86	0.93	0.96	0.98	0.99	1	0.96
<b>APCP(ttp)</b>	Time to <i>PC</i> of APCP	min	0.01	0.5	726	0.92	0.95	0.97	0.98	0.99	1	0.98
<b>Bud</b>	Time of bud emergence	min	0.01	0.5	913	0.77	0.91	0.95	0.97	0.99	1	0.95
<b>Cdc6</b>	<i>PC</i> of Cdc6	au	0.01	1e-4	948	0.82	0.89	0.95	0.97	0.99	1	0.96
<b>Cdc6(ttp)</b>	Time to <i>PC</i> of Cdc6	min	0.01	0.5	836	0.83	0.86	0.92	0.94	0.97	0.99	0.95
<b>Clb2</b>	<i>PC</i> of B-type cyclin Clb2	au	0.01	1e-4	903	0.84	0.92	0.96	0.98	0.99	1	0.97
<b>Clb2(ttp)</b>	Time to <i>PC</i> of Clb2	min	0.01	0.5	749	0.94	0.96	0.97	0.98	0.99	1	0.98
<b>Clb5</b>	<i>PC</i> of B-type cyclin Clb5	au	0.01	1e-4	935	0.90	0.95	0.98	0.98	0.99	1	0.98
<b>Clb5(ttp)</b>	Time to <i>PC</i> of Clb5	min	0.01	0.5	779	0.90	0.93	0.96	0.97	0.99	1	0.97
<b>Cln2</b>	<i>PC</i> of cyclin Cln2	au	0.01	1e-4	914	0.78	0.90	0.95	0.97	0.99	1	0.96
<b>Cln2(ttp)</b>	Time to <i>PC</i> of Cln2	min	0.01	0.5	790	0.90	0.94	0.96	0.98	0.99	1	0.98
<b>Rep</b>	Start time of DNA replication	min	0.01	0.5	932	0.91	0.93	0.96	0.98	0.99	1	0.97
<b>Sic1</b>	<i>PC</i> of cyclin-dep. kinase inhibitor Sic1	au	0.01	1e-4	950	0.77	0.89	0.94	0.97	0.99	1	0.95
<b>Sic1(ttp)</b>	Time to <i>PC</i> of Sic1	min	0.01	0.5	751	0.82	0.84	0.89	0.92	0.96	0.99	0.93
<b>Spn</b>	Time of completed chromosome alignment	min	0.01	0.5	816	0.94	0.96	0.97	0.98	0.99	1	0.98

**Table S9. Summary of phenotypic values, variability thresholds and distribution of  $V_A/V_G$  ratios for the circadian model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold. Abbreviations: phosphorylated – phos., cytosolic – cyt., nuclear – nuc., bottom concentration – b.c., peak concentration – p.c.

Phenotype	Description (see legend for abbreviations)	Units	Variability threshold		# of valid datasets	Quantiles and mean values of $V_A/V_G$						
			rel.	abs.		Q <sub>0.05</sub>	Q <sub>0.1</sub>	Q <sub>0.2</sub>	Q <sub>0.3</sub>	Q <sub>0.5</sub>	Q <sub>0.8</sub>	mean
<b>BC</b>	B.c. of cyt. BMAL1 protein	nM	0.01	1e-8	979	0.43	0.65	0.91	0.97	0.99	1	0.91
<b>BC(ttp)</b>	Time to p.c. of cyt. BMAL1 protein	hours	0.01	1e-8	997	0.40	0.70	0.85	0.91	0.96	0.98	0.88
<b>BCP</b>	B.c. of cyt. phos. BMAL1 protein	nM	0.01	1e-8	953	0.82	0.88	0.93	0.96	0.99	1	0.95
<b>BCP(ttp)</b>	Time to p.c. of cyt. phos. BMAL1 protein	hours	0.01	1e-8	999	0.62	0.80	0.87	0.92	0.97	0.99	0.91
<b>BN</b>	B.c. of nuc. BMAL1 protein	nM	0.01	1e-8	999	0.37	0.67	0.84	0.93	0.98	0.99	0.89
<b>BN(ttp)</b>	Time to p.c. of nuc. BMAL1 protein	hours	0.01	1e-8	1000	0.48	0.72	0.83	0.89	0.95	0.98	0.88
<b>BNP</b>	B.c. of nuc. Phos. BMAL1 protein	nM	0.01	1e-8	979	0.82	0.91	0.97	0.99	1	1	0.97
<b>BNP(ttp)</b>	Time to p.c. of nuc. Phos. BMAL1 protein	hours	0.01	1e-8	999	0.66	0.82	0.89	0.94	0.98	0.99	0.93
<b>CC</b>	B.c. of cyt. CRY protein	nM	0.01	1e-8	1000	0.59	0.63	0.69	0.74	0.82	0.95	0.81
<b>CC(ttp)</b>	Time to p.c. of cyt. CRY protein	hours	0.01	1e-8	1000	0.10	0.16	0.27	0.39	0.67	0.94	0.61
<b>CCP</b>	B.c. of cyt. phos. CRY protein	nM	0.01	1e-8	969	0.61	0.68	0.74	0.79	0.86	0.95	0.84
<b>CCP(ttp)</b>	Time to p.c. of cyt. phos. CRY protein	hours	0.01	1e-8	1000	0.10	0.15	0.31	0.46	0.70	0.91	0.62
<b>IN</b>	B.c. of inactive complex between CLOCK-BMAL1 and PER-CRY	nM	0.01	1e-8	1000	0.51	0.62	0.73	0.80	0.87	0.95	0.83
<b>IN(ttp)</b>	Time to p.c. of inactive complex between CLOCK-BMAL1 and PER-CRY	hours	0.01	1e-8	1000	0.22	0.36	0.53	0.66	0.82	0.95	0.73
<b>MB</b>	B.c. of mRNA of the <i>Bmall</i> gene	nM	0.01	1e-8	901	0.32	0.50	0.79	0.93	0.99	1	0.88
<b>MB(ttp)</b>	Time to p.c. of mRNA of the <i>Bmall</i> gene	hours	0.01	1e-8	894	0.19	0.53	0.78	0.82	0.88	0.95	0.81
<b>MC</b>	B.c. of mRNA of the <i>Cry</i> gene	nM	0.01	1e-8	999	0.91	0.92	0.94	0.96	0.98	0.99	0.96

<b>MC(ttp)</b>	Time to p.c. of mRNA of the <i>Cry</i> gene	hours	0.01	1e-8	1000	0.52	0.66	0.82	0.88	0.96	0.99	0.88
<b>MP</b>	B.c. of the mRNA of the <i>Per</i> gene	nM	0.01	1e-8	997	0.90	0.92	0.94	0.95	0.98	0.99	0.96
<b>MP(ttp)</b>	Time to p.c. of the mRNA of the <i>Per</i> gene	hours	0.01	1e-8	999	0.32	0.52	0.74	0.87	0.96	0.99	0.85
<b>PC</b>	B.c. of cyt. PER protein	nM	0.01	1e-8	1000	0.41	0.48	0.57	0.63	0.72	0.88	0.72
<b>PC(ttp)</b>	Time to p.c. of cyt. PER protein	hours	0.01	1e-8	1000	0.04	0.07	0.14	0.23	0.46	0.88	0.49
<b>PCC</b>	B.c. of cyt. PER-CRY complex	nM	0.01	1e-8	1000	0.8	0.83	0.87	0.90	0.93	0.97	0.92
<b>PCC(ttp)</b>	Time to p.c. of cyt. PER-CRY complex	hours	0.01	1e-8	1000	0.49	0.62	0.77	0.84	0.92	0.98	0.85
<b>PCCP</b>	B.c. of cyt. Phos. PER-CRY complex	nM	0.01	1e-8	1000	0.92	0.94	0.95	0.96	0.98	0.99	0.97
<b>PCCP(ttp)</b>	Time to p.c. of cyt. phos. PER-CRY complex	hours	0.01	1e-8	1000	0.46	0.61	0.75	0.83	0.92	0.98	0.85
<b>PCN</b>	B.c. of nuc. PER-CRY complex	nM	0.01	1e-8	1000	0.52	0.60	0.69	0.75	0.85	0.95	0.81
<b>PCN(ttp)</b>	Time to p.c. of nuc. PER-CRY complex	hours	0.01	1e-8	999	0.62	0.72	0.84	0.89	0.95	0.98	0.90
<b>PCNP</b>	B.c. of nuc. Phos. PER-CRY complex	nM	0.01	1e-8	1000	0.80	0.83	0.87	0.90	0.94	0.98	0.92
<b>PCNP(ttp)</b>	Time to p.c. of nuc. phos. PER-CRY complex	hours	0.01	1e-8	999	0.59	0.70	0.83	0.89	0.95	0.99	0.89
<b>PCP</b>	B.c. of cyt. phos. PER protein	nM	0.01	1e-8	989	0.70	0.75	0.81	0.85	0.91	0.96	0.88
<b>PCP(ttp)</b>	Time to p.c. of cyt. phos. PER protein	hours	0.01	1e-8	1000	0.04	0.07	0.15	0.24	0.46	0.88	0.50
<b>Period</b>	Time for one complete cycle	hours	0.01	0.1	997	0.60	0.81	0.90	0.94	0.98	0.99	0.92

**Table S10. Summary of phenotypic values, variability thresholds and distribution of  $V_A/V_G$  ratios for the action potential model.** The first three columns list the phenotype abbreviations used in this study, a text description of the phenotypes and their units. The thresholds used to filter out dataset with very low relative and/or absolute variability are listed in the next two columns, followed by the number of Monte Carlo simulations (out of 1000) passing the threshold. The last 7 columns contain quantiles and means of the  $V_A/V_G$  values for the datasets passing the variability threshold.

Phenotype	Description	Units	Variability threshold		# of valid datasets	Quantiles and mean values of $V_A/V_G$							
			rel.	abs.		Q <sub>0.05</sub>	Q <sub>0.1</sub>	Q <sub>0.2</sub>	Q <sub>0.3</sub>	Q <sub>0.5</sub>	Q <sub>0.8</sub>	mean	
Action Potential	25%	Time to 25% of the initial value	ms	0.01	1e-5	909	0.31	0.44	0.67	0.72	0.88	0.98	0.88
	50%	Time to 50% of the initial value	ms	0.01	1e-5	980	0.29	0.39	0.63	0.72	0.90	1	0.9
	75%	Time to 75% of the initial value	ms	0.01	1e-5	968	0.21	0.31	0.48	0.62	0.88	0.99	0.88
	90%	Time to 90% of the initial value	ms	0.01	1e-5	976	0.21	0.31	0.64	0.92	0.99	1	0.99
	Amplitude	Amplitude of action potential value	mV	0.01	1e-5	652	0.20	0.28	0.50	0.66	0.97	0.99	0.97
	Base	Initial action potential	mV	0.01	1e-5	1000	0.21	0.31	0.56	0.67	0.83	0.98	0.83
	Peak	Maximum action potential	mV	0.01	1e-5	892	0.19	0.27	0.45	0.59	0.77	0.99	0.77
	Time to peak	Time to maximum action potential	ms	0.01	1e-5	801	0.20	0.30	0.48	0.59	0.83	0.98	0.84
Calcium transient	25%	Time to 25% of the initial concentration	ms	0.01	1e-5	961	0.18	0.28	0.53	0.69	0.92	0.99	0.92
	50%	Time to 50% of the initial concentration	ms	0.01	1e-5	960	0.22	0.35	0.58	0.67	0.89	0.99	0.88
	75%	Time to 75% of the initial concentration	ms	0.01	1e-5	955	0.35	0.45	0.60	0.67	0.86	0.99	0.86
	90%	Time to 90% of the initial concentration	ms	0.01	1e-5	928	0.29	0.38	0.52	0.63	0.76	0.98	0.76
	Amplitude	Amplitude of the calcium concentration	uM	0.01	1e-5	990	0.18	0.29	0.46	0.58	0.77	0.98	0.77
	Base	Initial calcium concentration	uM	0.01	1e-5	965	0.27	0.36	0.51	0.63	0.78	0.99	0.78
	Peak	Maximum calcium concentration	uM	0.01	1e-5	983	0.28	0.36	0.52	0.61	0.73	0.98	0.73
	Time to peak	Time to maximum concentration	ms	0.01	1e-5	974	0.18	0.24	0.43	0.60	0.84	0.97	0.84

# On the relationship between heritability and regulatory architecture

Yunpeng Wang, Jon Olav Vik, Stig W. Omholt, Arne B. Gjuvsland\*

## Supplementary text

Here we describe the five models in some more detail than in the main text.

### cAMP Model

The complete cAMP signaling pathway [1] of *Saccharomyces cerevisiae*, involved in various essential cell activities such as nutrient sensing, stress response, growth, cell cycle progression, is modeled by a system of ordinary differential equations having 15 variables and 27 parameters. Glucose metabolism via the glycolysis pathway as well as the dynamics of the G-protein Gpa2a, the Kelch repeat homologue protein (Krh), the G-protein Ras2, and Protein Kinase A (PKA) is described by mass action kinetics. Modified Michaelis-Menten kinetics is used to describe the activity of adenylate cyclase and phosphodiesterase. A number of negative feedback mechanisms are included in the model. The model is capable of reproducing how cAMP changes in response to addition of glucose both in wild-type and several mutant strains.

### Glycolysis model

The branched kinetic model of the glycolysis pathway of non-growing anaerobic *Saccharomyces cerevisiae* [2] how external glucose molecules are transported into the cell, its conversion into pyruvate (ultimately to ethanol), glycogen, trehalose, glycerol, and succinate, by a series of enzyme-catalyzed reactions. The dynamics of 19 metabolic fluxes (12 reversible and 7 irreversible) and 17 intermediate concentrations, are represented by 14 ordinary differential equations. The model described the kinetics of the 13 enzymes hexokinase (HK), phosphoglucose isomerase (PGI), phosphofructokinase (PFK), aldolase (ALD), triosephosphate isomerase (TPI), glycerol 3-phosphate dehydrogenase (G3PDH), glyceraldehyde-3-phosphate dehydrogenase (GraPDH), phosphoglycerate kinase (PGK), phosphoglycerate mutase (PGM), enolase (ENO), pyruvate kinase (PYK) pyruvate decarboxylase (PDC) and alcohol dehydrogenase (ADH).

### **Cell cycle model**

A model based on the consensus mechanism of the cell cycle regulation in budding yeast [3], *Saccharomyces cerevisiae*. The model describes this process by 36 ordinary differential and 25 algebraic equations. The four phases in cell cycle, S (the primary activity is DNA synthesis), G2 (preparing for mitosis), M (mitosis) and G1 (growth phase) are modeled as two states, the G1 state and the S-G2-M state. The unidirectional transitions between the two self-maintaining states are achieved by a bistable mechanism involving positive and negative feedback loops. The model is capable of mimicking the behaviour of >100 mutant strains.

### **Circadian model**

The model describes a consensus network underlying the mammalian circadian clock [4,5], a regulatory network involving the *Per*, *Cry*, *Bmal1* and *Clock* genes. The model consists of 16 coupled ordinary differential equations, with intertwined positive and negative feedback loops. The state variable include mRNA, phosphorylated and unphosphorylated proteins as well as protein complexes. Few sleep-wake related human disorders were studied with this model by examining the parameter spaces. The model gives rise to sustained oscillation with a period close to 24 hours under continuous darkness.

### **Action potential model**

The mouse ventricular myocyte model [6] extends that of Bondarenko et al [7] with more realistic calcium handling, detailed re-parameterization to consistent experimental data for the C57BL/6 "black 6" mouse and conservation of charge. State variables include ion concentrations of sodium, potassium and calcium in the cytosol, calcium concentration in the sarcoplasmic reticulum, and the conformation state distribution of ion channels, whose transition rates between open, closed, and inactivated conformations may depend on transmembrane voltage. Formulated as a system of 35 coupled ordinary differential equations with more than hundred of parameters, this model provides a comprehensive representation of membrane-bound channels and transporter functions as well as fluxes between the cytosol and intracellular organelles.



1. Williamson T, Schwartz J-M, Kell DB, Stateva L (2009) Deterministic mathematical models of the cAMP pathway in *Saccharomyces cerevisiae*. *BMC Syst Biol* 3: 70. doi:10.1186/1752-0509-3-70.
2. Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, et al. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem* 267: 5313–5329.
3. Chen KC, Calzone L, Csikasz-Nagy A, Cross F, Novak B, et al. (2004) Integrative Analysis of Cell Cycle Control in Budding Yeast. *Molecular Biology of the Cell* 15: 3841–3862. doi:10.1091/mbc.E03-11-0794.
4. Leloup J-C, Goldbeter A (2003) Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci USA* 100: 7051–7056. doi:10.1073/pnas.1132112100.
5. Leloup J-C, Goldbeter A (2004) Modeling the mammalian circadian clock: sensitivity analysis and multiplicity of oscillatory mechanisms. *J Theor Biol* 230: 541–562. doi:10.1016/j.jtbi.2004.04.040.
6. Li L, Niederer SA, Idigo W, Zhang YH, Swietach P, et al. (2010) A mathematical model of the murine ventricular myocyte: a data-driven biophysically based approach applied to mice overexpressing the canine NCX isoform. *Am J Physiol Heart Circ Physiol* 299: H1045–H1063. doi:10.1152/ajpheart.00219.2010.
7. Bondarenko VE, Szigeti GP, Bett GCL, Kim S-J, Rasmusson RL (2004) Computer model of action potential of mouse ventricular myocytes. *Am J Physiol Heart Circ Physiol* 287: H1378–H1403. doi:10.1152/ajpheart.00185.2003.



# **Paper II**



## **Monotonicity is a key measure of genotype-phenotype maps**

Arne B. Gjuvsland<sup>1,4</sup>, Yunpeng Wang<sup>2</sup>, Erik Plahte<sup>1</sup> and Stig W. Omholt<sup>2,3</sup>

<sup>1</sup>Centre for Integrative Genetics (CIGENE), Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway, <sup>2</sup>Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Ås, Norway, <sup>3</sup>NTNU Norwegian University of Science and Technology, Department of Biology, Centre for Dynamics of Biological Diversity, Realfagsbygget, NO-7491 Trondheim, Norway. <sup>4</sup>To whom correspondence should be addressed. E-mail: arne.gjuvsland@umb.no

*Running title:* Monotonicity of genotype-phenotype maps

*Subject categories:* Simulation and data analysis / Metabolic and regulatory networks

*Total character count (excluding the Methods section, tables and supplementary material):* 26950

## **Abstract**

It was recently shown that monotone gene action, i.e. the order-preservation between allele content and corresponding phenotypic 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 hypothesis that 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 and positive feedback, as well as pleiotropy in the mapping between genotypes and gene regulatory parameters, are clearly predisposed for generating non-monotonicity. These deep connections between molecular regulatory architecture and properties of the genotype-phenotype map do not materialize from the classical distinction between additive and nonadditive gene action.

Keywords:

genotype-phenotype map / gene regulatory networks / genetic variance / monotonicity

## Introduction

Francis Galton's letter to Karl Pearson in 1897 describes quite well the development of population and quantitative genetics theory the last 100 years: "... the solutions to these problems [variation, heredity, selection, and other phenomena relating to evolution] are in the first place statistical, the second place statistical, and only in the third place biological" (Pearson, 1914). To a high degree this statement also applies to recent systems genetics work based on high-dimensional -omics data (Mackay et al, 2009; Nadeau & Dudley, 2011). More specifically, quantitative genetics is based on orthogonal decomposition of the genotype-phenotype relation (GP map) and corresponding genotype frequencies into an additive part, and two types of non-additive parts, dominance and epistasis (Lynch & Walsh, 1998). Although this provides the basis for a very successful theory when it comes to predicting selection response and breeding values, it does not bring with it an understanding of what ensures the ubiquitous presence of high additive variance relative to genetic variance in sexually reproducing organisms across the whole allele frequency spectrum (Gjuvslund et al, 2011; Hill et al, 2008). The reason for this is that current genetic theory is built upon the empirical fact that offspring resemble their parents, and it is therefore not in a position to explain the biological foundation of this fundamental aspect of heredity. New approaches are thus needed to understand the mechanisms underlying this prerequisite for natural selection to operate.

It was recently shown (Gjuvslund et al, 2011) 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. By developing new quantitative measures for monotonicity we show that the GP maps arising from a wide range of gene regulatory network motifs are in general highly monotone. But most noteworthy, these new measures also reveal which motifs predispose for non-monotonicity. We show that these deep connections 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 (Lynch & Walsh, 1998).

Our approach provides a theoretical foundation for characterizing the degree of monotonicity of the genotype-phenotype mappings emerging from causally

cohesive genotype-phenotype models (Rajasingh et al, 2008; Vik et al, 2011) at any level of model complexity. This opens for a systematic study of the order-preserving and order-breaking properties of molecular regulatory structures underlying the whole spectrum of physiological regulation.

## Results

### 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 (alleles indexed 1 and 2) loci 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  $g_{kl} = g_k g_l \in \{1111, 1112, 1122, 1211, \dots, 2212, 2222\}$ . The general  $N$  loci genotype space  $\Gamma$  contains  $3^N$  genotypes  $g_1 g_2 \dots g_N$  (in condensed notation  $g_{1:N}$ ) constructed by concatenating elements from single locus genotype spaces,  $\Gamma = \{g_1 g_2 \dots g_N \mid g_k \in \{11, 12, 22\}, k = 1, 2, \dots, N\}$ . We use the allele content (i.e. the number of 2-alleles) of genotypes to define a partial order on the genotype space  $\Gamma$  (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  $\Gamma$ .

A genotype-phenotype map is a mapping  $G$  which assigns to each genotype  $g \in \Gamma$  a real-valued genotypic value (the mean trait value for a given genotype)  $G(g)$ . We define monotonicity of  $G$  in terms of how it transforms the partial genotype orders  $g$  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  $G(1111 \dots 11)$  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

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



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 examples of monotone and non-monotone GP maps including orthogonal components used to decompose genetic variance (Cheverud & Routman, 1996; Zeng et al, 2005), and classical dominance and epistasis patterns.

## 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 of 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 how close  $G$  is to the closest fully monotone GP map. In Supplementary Text 1 we provide a step-by-step example on measuring monotonicity of a real-world GP map from (Cheverud & Routman, 1995).

### Measure 1: quantifying non-monotonicity by substitution effects

A straightforward way to check whether the inequality (2) is fulfilled is to look at the signs of the effect of substituting a single allele at locus  $k$

$$s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N}) = G(\mathbf{g}_{1:k-1}i2\mathbf{g}_{k+1:N}) - G(\mathbf{g}_{1:k-1}i1\mathbf{g}_{k+1:N}) \quad (3)$$

while keeping the other allele ( $i = 1$  or  $2$ ) at locus  $k$  and the background genotype  $\mathbf{g}_{1:k+1}\mathbf{g}_{k+1:N}$  fixed. Monotonicity as defined above is equivalent to  $s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N})$  being non-negative across all genetic backgrounds of locus  $k$ . By taking into account also the magnitude of the substitution effects we can quantify monotonicity. We start with the set  $S^k = \{s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N})\}$  of single allele substitution effects for locus  $k$  for  $i=1,2$  and across all genotypic backgrounds  $\mathbf{g}_{1:k+1}\mathbf{g}_{k+1:N}$ . 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;  $S^k_+ = \{s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N}) \in S^k, s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N}) > 0\}$  and  $S^k_- = \{s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N}) \in S^k, s^i(\mathbf{g}_{1:k-1}\mathbf{g}_{k+1:N}) < 0\}$ . We compute the sum of positive substitution effects and the sum of absolute values of negative substitution effects,

$$\begin{aligned}
P_k &= \sum_{S_+^k} s^i(\mathbf{g}_{1:k-1} \mathbf{g}_{k+1:N}), \\
N_k &= \sum_{S_-^k} |s^i(\mathbf{g}_{1:k-1} \mathbf{g}_{k+1:N})|,
\end{aligned} \tag{4}$$

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

$$m_k = \frac{|P_k - N_k|}{T_k}. \tag{5}$$

Clearly  $G$  is monotone w.r.t. locus  $k$  iff  $m_k = 1$ , 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) as well as the GP maps showing partial dominance and duplicate dominance epistasis, whereas  $m_1 = m_2 = 0$  for the maps with pure overdominance (OD) and pure epistasis (AxA and DxD).

In order to describe the overall monotonicity of the GP map  $G$  we introduce the *degree of monotonicity* ( $m$ ) which is a weighted mean of the  $m_k$ , where the weights reflect the relative importance of the loci in terms of  $T_k$ ,

$$m = \frac{\sum_{k=1}^N m_k T_k}{\sum_{k=1}^N T_k}. \tag{6}$$

As shown in Figure 2A, the *degree of monotonicity* is 1 for the monotone maps in Figure 1 while it is 0 for the pure overdominance and pure epistasis maps. This definition of *degree of monotonicity* allows us to establish a vocabulary which is analogous to the classification of single locus dominance; a GP map is called *monotone* if  $m = 1$ , (*partially*) *non-monotone* if  $m < 1$  and *purely non-monotone* if  $m = 0$ . A more detailed presentation of the degree of monotonicity is given in section 1 of Supplementary Text 2.

The simulation results for random GP maps presented in Figure 2A show that there is a strong positive correlation between the degree of monotonicity and the size of the additive component ( $V_A^{UWR} / V_G^{UWR}$ ), both for the GP maps depicted in Figure 1

and for randomly sampled GP maps. A similar relationship was observed for three-locus random GP maps (Supplementary Figure 1A). All GP maps in Figure 2B with  $m < 0.2$  have  $V_A^{UWR} / V_G^{UWR} < 0.1$ . 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  $V_A^{UWR} / V_G^{UWR}$  as low as 0.375.

## Measure 2: quantifying monotonicity by isotonic regression

This measure quantifies the monotonicity by means of the distance between  $G$  and the closest monotone map. Consider a space of genotype-phenotype maps for  $N$  biallelic loci  $T = \{H \mid H : \Gamma \rightarrow R\}$  and define the subset  $T_M \subset T$  of monotone GP maps  $T_M = \{H \in T \mid H \text{ is order-preserving on } \Gamma\}$ . Given a GP map  $G \in T$  we identify the *monotone component* of  $G$  as the map  $G_M \in T_M$  which minimizes the residual variance  $\text{var}(G - G_M)$ , i.e.  $G_M$  is the monotone GP map which is closest to  $G$  in the least-squares sense. The monotone component  $G_M$  can be identified numerically by isotonic regression (Leeuw et al, 2009) of  $G$  subject to the partial ordering of genotypes defined in eq. (1). Furthermore the residual  $G_N$  is orthogonal to  $G_M$  in the sense that the vector product  $G_M \cdot G_N = 0$ . This allows the orthogonal decomposition

$$G = G_M + G_N, \quad (7)$$

of a genotype-phenotype map into a *monotone component*  $G_M$  and a *non-monotone component*  $G_N$  such that  $\text{var}(G) = \text{var}(G_M) + \text{var}(G_N)$ . Supplementary Text 1 illustrates the monotone and non-monotone components of the GP map reported in (Cheverud & Routman, 1995). The orthogonality property allows us to measure monotonicity of  $G$  in terms of the coefficient of determination  $R_{mono}^2$  of the isotonic regression given by the ratio  $R_{mono}^2 = \text{var}(G_M) / \text{var}(G)$ . In the case that  $G$  itself is monotone for all loci we have  $R_{mono}^2 = 1$ , while order-breaking for one or more loci will result in  $R_{mono}^2 < 1$ .

Simulation results for random GP maps show that  $R_{mono}^2$  is positively correlated to the size of the additive component (Figure 2B for two-locus GPs maps and Supplementary Figure 1B for three-locus GP maps) and that for a given  $V_A^{UWR} / V_G^{UWR}$  the lower bound for  $R_{mono}^2$  is close to a straight line from (0, 0.2) to (1, 1). However,

due to the search for the closest monotone GP map,  $R_{mono}^2$  will not become zero even for purely overdominant or purely epistatic maps. Thus measuring monotonicity by isotonic regression works very well for GP maps that are close to monotone, but distinguishes poorly between highly non-monotone maps. As shown in Supplementary Figure 2, the two monotonicity measures are highly correlated.

### **An R package for studying monotonicity in GP maps**

We developed an R package `gpmmap` for studying functional properties of genotype-phenotype maps. The package takes genotype-phenotype maps in the form of vectors of genotypic values as input, and provides (i) functions for determining whether the map is order-breaking or order-preserving w.r.t. any given locus, (ii) the degree of monotonicity  $m$ , (iii)  $R_{mono}^2$  using isotonic regression from the `isotone` package (Leeuw et al, 2009), and (iv) plots of the original and decomposed GP maps. The package is available<sup>1</sup> from CRAN under GPLv3.

### **Monotonicity in GP maps arising from gene regulatory networks**

Having established the measures of monotonicity in generic GP maps, we moved on to study the GP maps emerging from gene regulatory network models. We generated 1881 networks motifs of three genes  $X_1$ ,  $X_2$  and  $X_3$ , each with 0, 1 or 2 transcriptional regulators. For each network we generated 1000 three-locus GP maps by sampling allelic variation on maximal production rates and parameters in the gene regulation functions, and then recording the steady state expression level of  $X_3$  as the phenotype. (See Models and Methods for detailed descriptions of network connectivity, ODEs describing the dynamics of gene expression and Monte Carlo simulation procedures). Based on earlier results (Gjuvsland et al, 2007a; Gjuvsland et al, 2011) we hypothesized that incoherent feed forward or positive feedback 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 I 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

---

<sup>1</sup> *Note to reviewers: For review purposes the source code and manual is available at <http://arken.umb.no/~arnegi/R.gpmmap/>. The package will be published on <http://cran.r-project.org/> upon acceptance of the manuscript.*

Methods for details). Motifs with less than 100 usable GP maps were discarded from further analysis. For the genotype-to-parameter maps without pleiotropy (one locus determines one parameter, see Model and Methods) 868 motifs were discarded, while for the genotype-to-parameter map with pleiotropy (each locus determines three parameters) 791 motifs were discarded. All (but one) discarded motifs contained at least one positive feedback loop (Table I). 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.

We analyzed the monotonicity properties of 1090 and 1013 motifs for genotype-to-parameter maps with and without pleiotropy, respectively. The introduction of pleiotropy in the genotype to parameter map has a marked effect on the monotonicity characteristics of the associated GP map (Figure 3). With no pleiotropy the GP maps come out as highly monotone with a large majority being fully monotone or order-breaking for just a single locus. With pleiotropy, the majority of GP maps still show order-breaking either for no loci or just one locus, but a considerable number of GP maps are order-breaking for two or three loci. Furthermore, dividing the motifs into the four groups given in Table I, 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 appear 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 (Figure 3A). In Supplementary Text 2 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 feed-forward 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 mapping increases the frequency of order-breaking GP maps substantially (Figure 3B). Motifs lacking both incoherent feedforward and positive feedback may now 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 4). We also find (Figure 4) that for all classes of motifs the majority of GP maps are fully monotone, while the most non-monotone GP maps (lowest  $R_{mono}^2$  values) are observed for motifs with positive feedback.

## 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 & Mackay, 1996; Lynch & 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 2) is expected.

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. (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. (Wu et al, 2009) developed a method to screen for and test the significance of essential interaction in genome-wide association studies.

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, because monotonicity as we have defined it, is a mathematical concept, our approach 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 & Routman, 1995) or functional (Hansen & Wagner, 2001) descriptor rather than a statistical one. The distinction between physiological and statistical epistasis has led to much debate (Phillips, 2008). Zeng et al. (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 separate from any particular study population is essential if we want to connect quantitative genetics with regulatory biology. Thus there is clearly both an epistemic and an instrumental dimension attached to the concept of monotonicity of GP maps.

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 finding (Gjuvsland et al, 2007a) that gene regulatory motifs with positive feedback give considerably more statistical epistasis than those without.

Without any restrictions on the connectivity of a three-gene system there are  $3^9 = 19683$  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 (Gjuvsland et al, 2007b; Plahte et al, 1998). Model formalisms allowing an arbitrary number of regulators are well established (Siegal & Bergman, 2002; Wagner, 1994; Wagner, 1996) and could be extended to diploid forms and used in later studies.

Even though both incoherent feedforward and positive feedback predispose for non-monotone GP maps, the underlying mechanisms are somewhat 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). In networks with positive feedback differences in loop gain between genotypes can give non-monotonicity (see Supplementary Text 2 for details). Positive feedback also predisposes for multiple steady states, and order-breaking could 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 loop in the connectivity matrix  $A$  of a system does not guarantee that the loop is active at any point during the time course of the system. This could explain the pattern in Figure 3 where motifs with only positive feedback lead to order-breaking far less frequently than those that also contain incoherent feedforward.

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 comprehensive 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.

## Materials and Methods

### Differential equation models

We study a family of 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 of or both the other genes. The wiring of the system is described by a 3x3 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 gene  $l$  is not a regulator of gene  $k$ , if  $A_{kl} = 1$  then gene  $l$  is an activator of gene  $k$  and if  $A_{kl} = -1$  then gene  $l$  is a repressor of gene  $k$ . 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 S3. We use 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,

$$\begin{aligned}\dot{x}_{k1} &= \alpha_{k1} R_{k1}(y_1, y_2, y_3) - \gamma_{k1} x_{k1}, \\ \dot{x}_{k2} &= \alpha_{k2} R_{k2}(y_1, y_2, y_3) - \gamma_{k2} x_{k2}, \\ y_k &= x_{k1} + x_{k2}, \quad k = 1, 2, 3.\end{aligned}\tag{8}$$

Here  $\alpha_{ki}$  is the maximal production rate for allele  $i$  of gene  $k$ ,  $\gamma_{ki}$  is the decay rate, while  $R_{ki}$  is the gene regulation function (dose-response function). If gene  $k$  has no regulators, we assume production is always switched on so  $R_{ki} = 1$ . If gene  $k$  has a single regulator  $y_l$ , 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 gene  $l$  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  $\theta_{lki}$  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  $y_l$  and  $y_j$  we set  $R_{ki}(y_l, y_j) = S(y_l, \theta_{lki}, p_{lki}) S(y_j, \theta_{jki}, p_{jki})$ .

With up to two regulators per gene the number of possible connectivity matrices is  $19^3 = 6859$ . We further require that the system is connected, and that gene 3 is not purely upstream of gene 1 and 2 so either gene 1 and 2 both regulate gene 3 directly ( $A_{31} A_{32} \neq 0$ ), or one regulates in 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 identify pairs of matrices which are symmetric with respect to interchanging  $X_1$  and  $X_2$  and pick just one matrix from each pair. The resulting 1881 connectivity matrices are used for our gene regulatory simulations.

## Identifying feedback loops and feedforward motifs

For a system defined by a given connectivity matrix we compute the loop products of three autoregulatory feedback loops  $FL_1 = A_{11}$ ,  $FL_2 = A_{22}$ ,  $FL_3 = A_{33}$ , three two-gene feedback loops:  $FL_{12} = A_{21} A_{12}$ ,  $FL_{13} = A_{31} A_{13}$ ,  $FL_{23} = A_{23} A_{32}$  and two three-gene feedback loops:  $FL_{123} = A_{32} A_{21} A_{13}$ ,  $FL_{213} = A_{31} A_{12} A_{23}$ . Nonzero 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 compute the products for two feedforward motifs:  $FFL_{32} = A_{32} (A_{31} A_{12})$ ,  $FFL_{31} = A_{31} (A_{32} A_{21})$ . Again nonzero 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. Supplementary Figure 3 depicts the connectivity matrix and the signed digraphs of a system with a positive feedback loop as well as a system with incoherent feedforward. Supplementary Dataset 1 contains adjacency matrices and loop products for all 1881 motifs.

## Genotype-phenotype simulations

The simulations were performed in a similar fashion as described in (Gjuvslund et al, 2011). We compared two different types of genotype-to-parameter maps:

- *Genotype to parameter map without pleiotropy*: biallelic genotypic variation for all three loci were introduced through the maximal production rates  $\alpha_{ki}$ , 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  $\alpha_{ki}$  (sampled from  $U(100, 200)$ ), regulation thresholds  $\theta_{lki}$  (sampled from  $U(20, 40)$ ), and regulation steepnesses  $p_{lki}$  (sampled from  $U(1, 10)$ ).

All decay rates  $\gamma_{ki}$  were set to 10. We assembled parameter sets for all 27 diploid genotypes, and for each genotypic parameter set the system of equations (8) 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 done on GP maps where the absolute range (maximum genotypic value – minimum genotypic value) and relative range (absolute range / mean genotypic value) were both  $>0.01$ .

## Statistical decomposition of genotype-phenotype maps

Statistical decomposition of GP maps was performed with the function `linearGPmapanalysis` in the R package `noia` (Le Rouzic & Alvarez-Castro, 2008) (<http://cran.r-project.org/web/packages/noia/>) version 0.94.1. We used unweighted regression (UWR) as in (Cheverud & Routman, 1995) to decompose a GP map into its additive and non-additive components and computed the ratio of additive to total genetic variance  $V_{UWR}^A / V_{UWR}^G$  as a measure of additivity. See (Alvarez-Castro & Carlborg, 2007; Zeng et al, 2005) for details on the theoretical framework.

## Acknowledgements

This work was supported by the Research Council of Norway (<http://www.rcn.no/>) under the eVITA program, project number 178901/V30 and by the Virtual Physiological Rat Project funded through NIH grant P50-GM094503. NOTUR, the Norwegian metacenter for computational science, provided computing resources under project nn4653k.

## Author Contributions

Conceived the study: ABG and SWO. Performed simulations and analysis: ABG, EP, and YW. Wrote the paper: all authors.

## Conflict of Interest

The authors declare that they have no conflict of interest.

## References

- Alvarez-Castro JM, Carlborg Ö (2007) A unified model for functional and statistical epistasis and its application in QTL analysis. *Genetics* **176**: 1151-1167
- Cheverud JM, Routman EJ (1995) Epistasis and its contribution to genetic variance components. *Genetics* **139**: 1455-1461
- Cheverud JM, Routman EJ (1996) Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution* **50**: 1042-1051

- Cockerham CC (1954) An extension of the concept of partitioning hereditary variances for analysis of covariances among relatives when epistasis is present. *Genetics* **39**: 859-882
- Falconer DS, Mackay TFC (1996) *Introduction to quantitative genetics*, Harlow: Longman Group.
- Fisher RA (1918) The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399-433
- Gjuvslund AB, Hayes BJ, Omholt SW, Carlborg Ö (2007a) Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* **175**: 411-420
- Gjuvslund AB, Plahte E, Omholt SW (2007b) Threshold-dominated regulation hides genetic variation in gene expression networks. *BMC Systems Biology* **1**: 57
- Gjuvslund AB, Vik JO, Woolliams JA, Omholt SW (2011) Order-preserving principles underlying genotype-phenotype maps ensure high additive proportions of genetic variance. *Journal of Evolutionary Biology* **24**: 2269-2279
- Hansen TF, Wagner GP (2001) Modeling genetic architecture: a multilinear theory of gene interaction. *Theoretical Population Biology* **59**: 61-86
- Hill WG, Goddard ME, Visscher PM (2008) Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genetics* **4**: e1000008
- Kaplan S, Bren A, Dekel E, Alon U (2008) The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Molecular Systems Biology* **4**: 203
- Le Rouzic A, Alvarez-Castro JM (2008) Estimation of genetic effects and genotype-phenotype maps. *Evolutionary Bioinformatics* **4**: 225-235
- Leeuw JD, Hornik K, Mair P (2009) Isotone Optimization in R : Pool-Adjacent-Violators Algorithm ( PAVA ) and Active Set Methods. *Journal of Statistical Software* **32**
- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*, Sunderland, Mass.: Sinauer Associates.
- Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**: 565-577
- Mestl T, Plahte E, Omholt SW (1995) A mathematical framework for describing and analysing gene regulatory networks. *Journal of Theoretical Biology* **176**: 291-300
- Nadeau JH, Dudley AM (2011) Systems Genetics. *Science* **331**: 1015-1016
- Omholt SW, Plahte E, Øyehaug L, Xiang KF (2000) Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* **155**: 969-980

- Pearson K (1914) *The life, letters and labours of Francis Galton*, Vol. IIIa, p. 1, Cambridge, UK: University press.
- Phillips PC (1998) The language of gene interaction. *Genetics* **149**: 1167-1171
- Phillips PC (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**: 855-867
- Plahte E, Mestl T, Omholt SW (1995) Feedback Loops, Stability and Multistationarity in Dynamical Systems. *Journal of Biological Systems* **3**: 409-413
- Plahte E, Mestl T, Omholt SW (1998) A methodological basis for description and analysis of systems with complex switch-like interactions. *Journal of Mathematical Biology* **36**: 321-348
- Rajasingh H, Gjuvsland AB, Våge DI, Omholt SW (2008) When parameters in dynamic models become phenotypes: a case study on flesh pigmentation in the Chinook salmon (*Oncorhynchus tshawytscha*). *Genetics* **179**: 1113-1118
- Siegal ML, Bergman A (2002) Waddington's canalization revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 10528-10532
- Vik JO, Gjuvsland AB, Li L, Tøndel K, Niederer S, Smith N, Hunter P, Omholt SW (2011) Genotype-phenotype map characteristics of an in silico heart cell. *Frontiers in Physiology* **2**
- Wagner A (1994) Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 4387-4391
- Wagner A (1996) Does evolutionary plasticity evolve? *Evolution* **50**: 1008-1023
- Wang X, Elston RC, Zhu X (2010) The meaning of interaction. *Human heredity* **70**: 269-277
- Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**: 1165-1174
- Wu C, Zhang H, Liu X, Dewan A, Dubrow R, Ying Z, Yang Y, Hoh J (2009) Detecting essential and removable interactions in genome-wide association studies. *Statistics and its interface* **2**: 161-170
- Zeng ZB, Wang T, Zou W (2005) Modeling quantitative trait loci and interpretation of models. *Genetics* **169**: 1711-1725

## Figure legends

**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 (colour) versus 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 I 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 (AxA) and dominance-by-dominance epistasis (DxD). The rightmost plot shows a GP map that is monotone w.r.t. locus 1, but non-monotone w.r.t. locus 2.

**Figure 2 Measures of monotonicity versus additivity of GP maps.** Scatterplots showing  $V_A/V_G$  from unweighted regression versus **(A)** degree of monotonicity and **(B)** monotone  $R^2$ . Black dots correspond to the maps shown in Figure 1 together with additive-by-dominance epistasis (AxD), a map with two loci showing complete dominance (CD) and two classical epistasis types from Table I 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 sorting to introduce order-preservation for 1 locus (see (Gjuvslund et al, 2011)).

**Figure 3 Order-breaking in motifs containing a single feedforward loop.**

Summary of order-breaking for all motifs for which there were at least 100 (out of 1000) valid Monte Carlo simulations. 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**). Colours indicate classes of motifs based on the presence/absence of incoherent feedforward and positive feedback loops, see Table I for the number of motifs in each class. A single boxplot summarizes, for all motifs in the given class, the proportion of the valid GP maps (y-axis) that are order-breaking with respect to a given number of loci (x-axis).

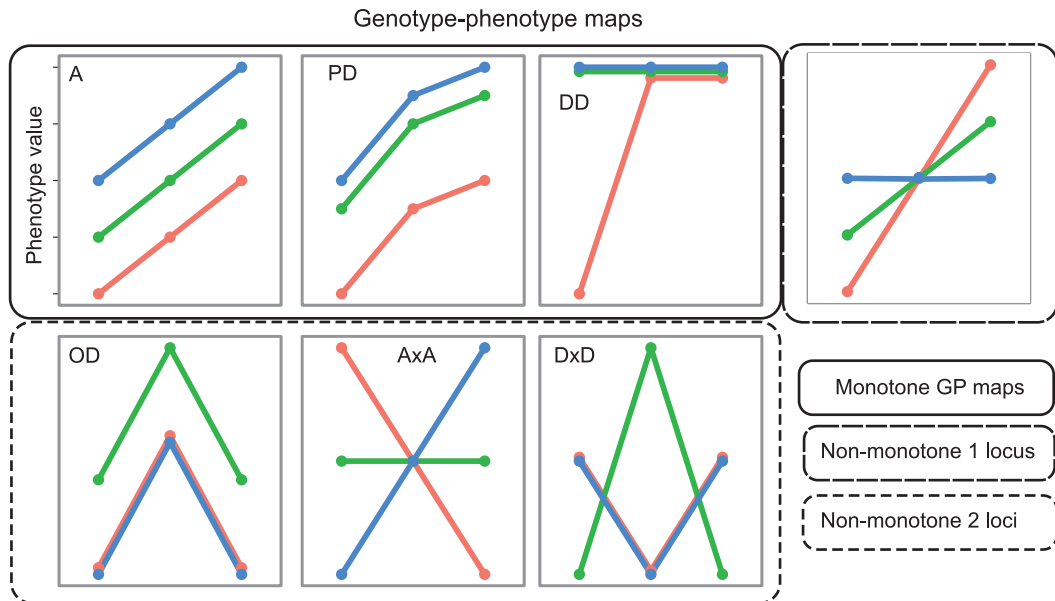
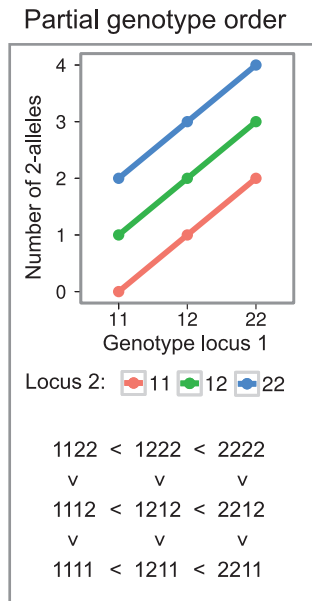
**Figure 4 Empirical distribution functions for monotone  $R^2$**  Summary of  $R^2$  values from isotone regression for all motifs for which there were at least 100 (out of 1000) valid Monte Carlo simulations. Results are shown for 1013 motifs with a genotype-to-parameter map without pleiotropy (**A**) 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 I for the number of motifs in each class. Each curve shows, for a single motif, the empirical distribution function value (y-axis) of monotone  $R^2$  for all valid genotype-phenotype maps (x-axis).

## Tables

<b>Table I: Frequencies (proportion of row total in parenthesis) of incoherent feedforward and positive feedback loops in subsets of the 1881 studied motifs.</b>					
<i>Dataset</i>	<i>Number of motifs</i>	<i>Motifs containing (proportion of)</i>			
		<i>Incoh. feedforward</i>		<i>No incoh. feedforward</i>	
		<i>Positive feedback</i>	<i>No positive feedback</i>	<i>Positive feedback</i>	<i>No positive feedback</i>
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)
Analysed 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
Analysed motifs	1090	163 (0.149)	48 (0.044)	627 (0.575)	252 (0.231)

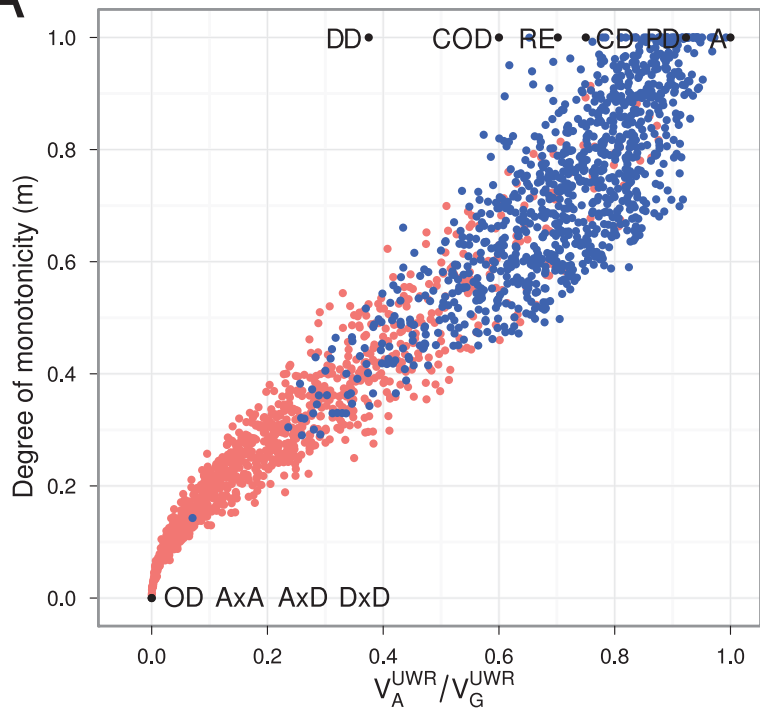


# Figure 1



# Figure 2

## A



## B

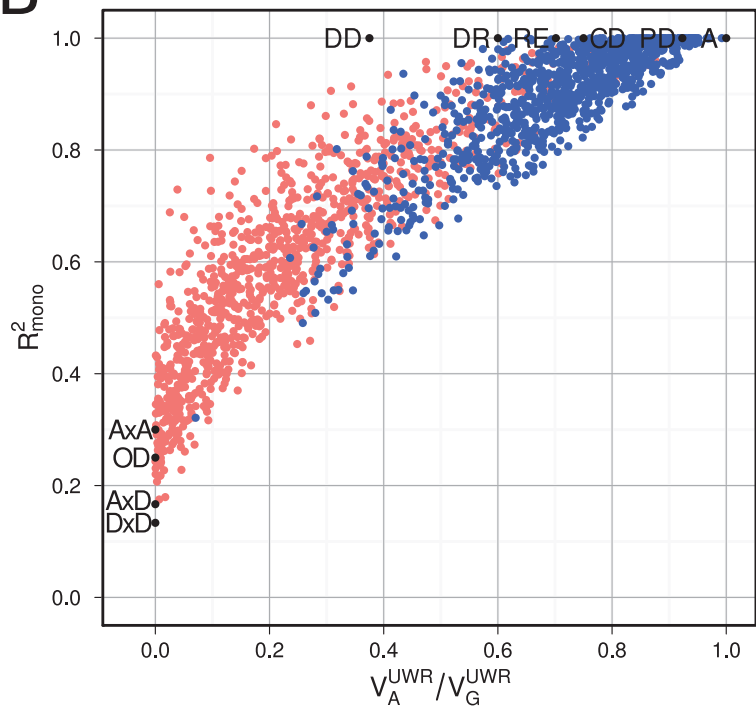


Figure 3

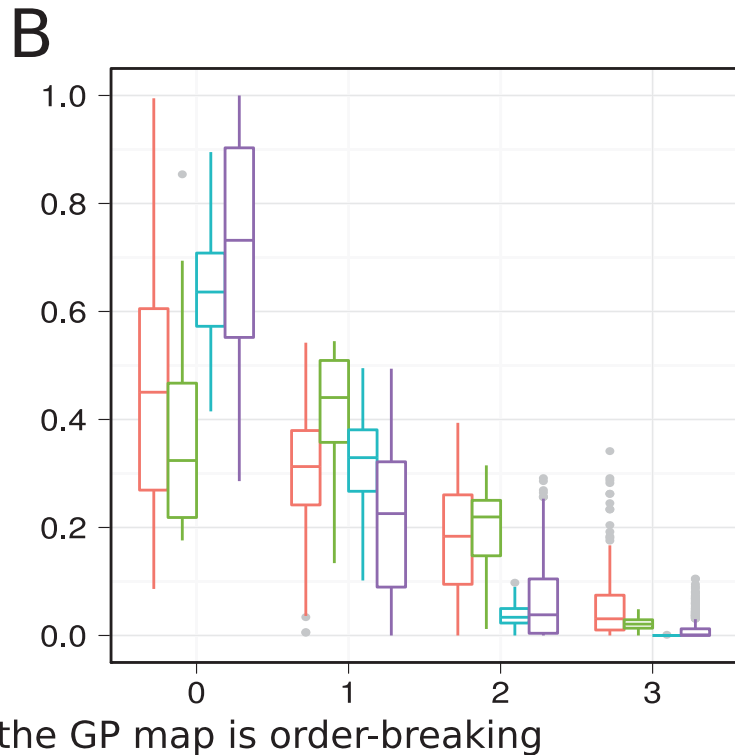
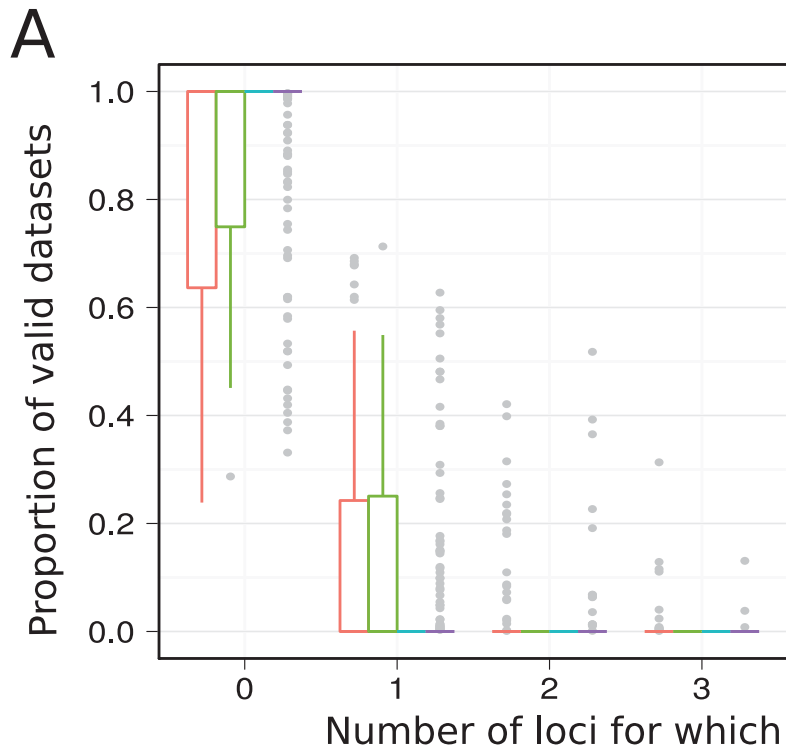
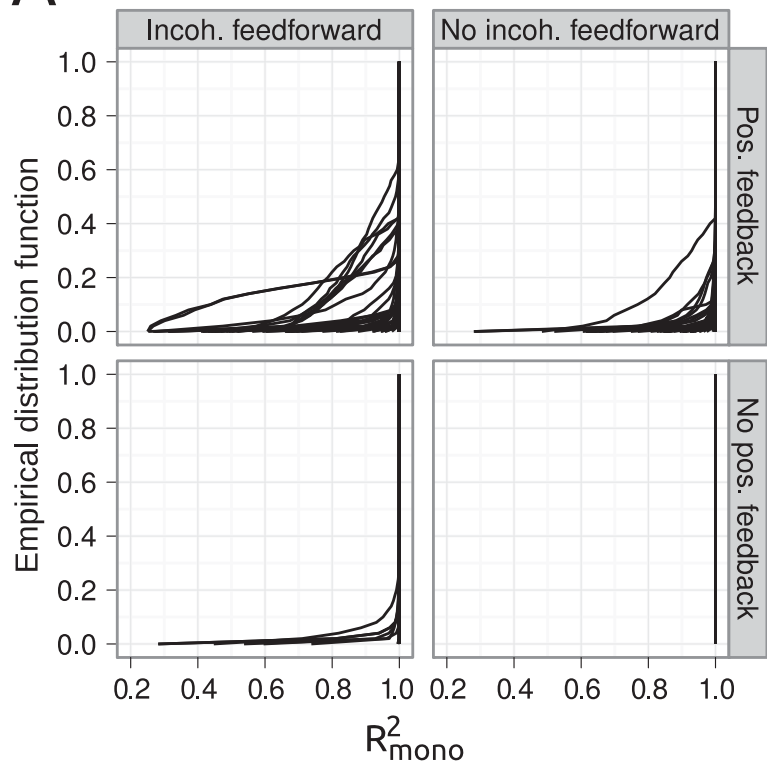
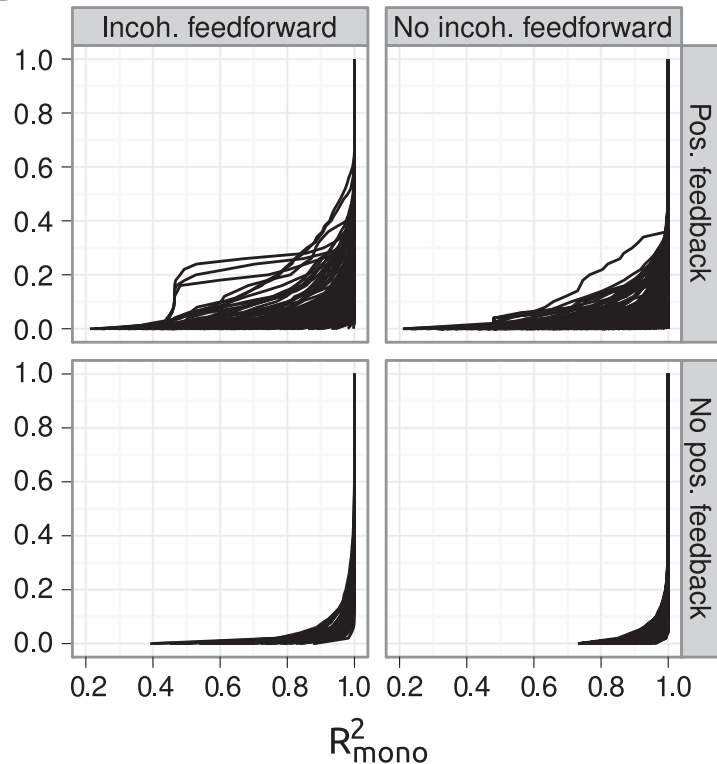


Figure 4

A



B



# Supplementary information

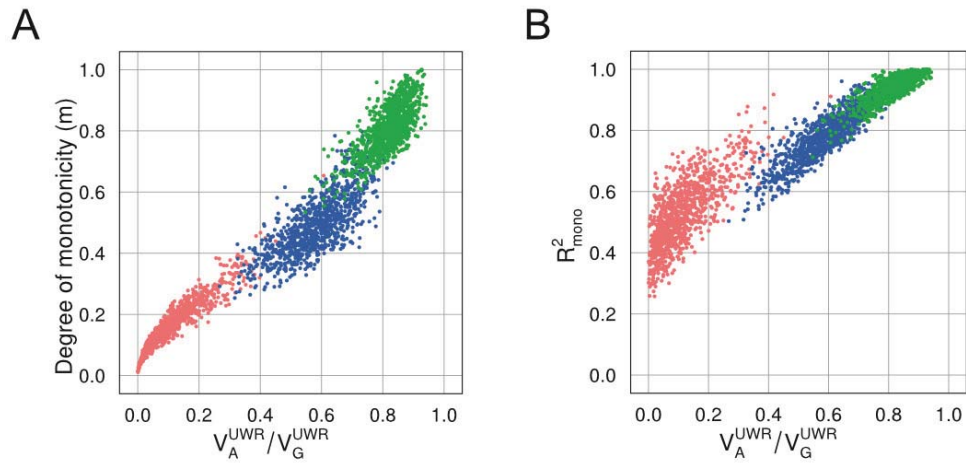
This document contains supplementary information for the manuscript

“Monotonicity is a key measure of genotype-phenotype maps”

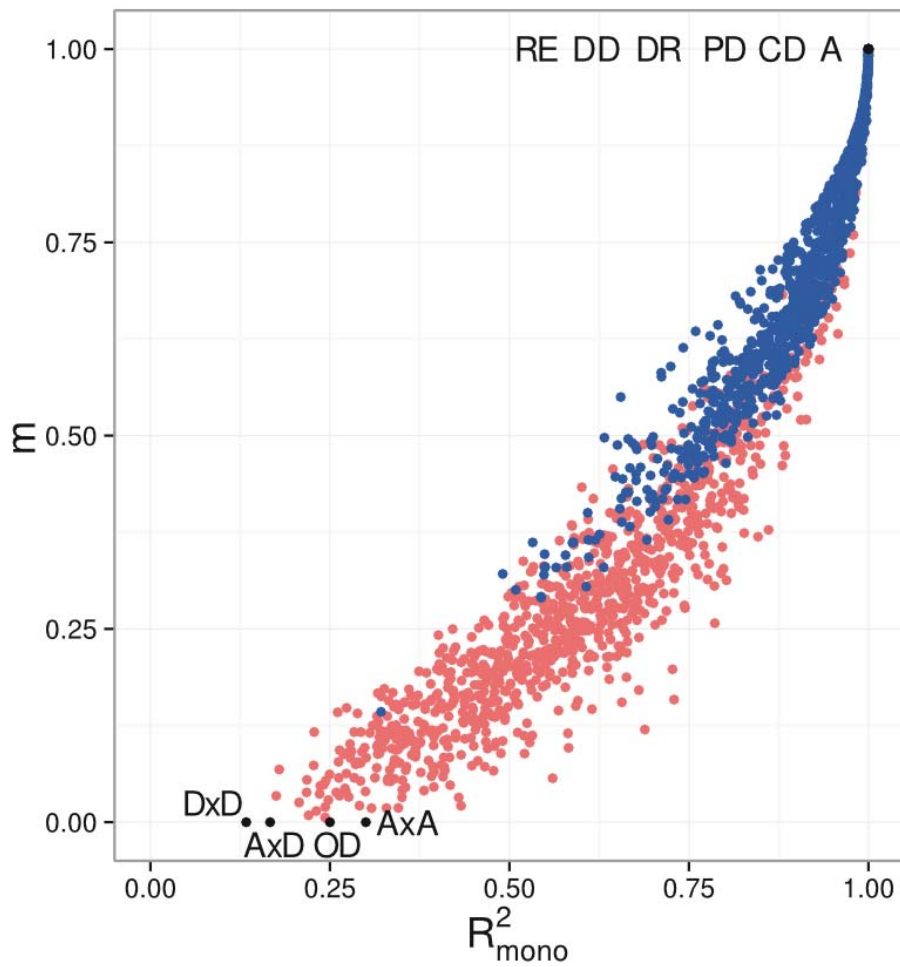
Arne B. Gjuvsland, Yunpeng Wang, Erik Plahte and Stig W. Omholt

## Table of Contents

- **Supplementary Figure 1** Measures of monotonicity versus additivity of GP maps with three loci.
- **Supplementary Figure 2** Comparing measures of monotonicity GP maps.
- **Supplementary Figure 3** Connectivity matrices and signed directed graphs.
- **Supplementary Text 1** Supplementary text (including **Supplementary Table 1** and **Supplementary Figure 4-5**) providing a detailed example on measuring monotonicity of a GP map.
- **Supplementary Text 2** Supplementary text providing analytical results on the relationship between feedforward/feedback in gene regulatory networks and order-breaking in the genotype-phenotype map.

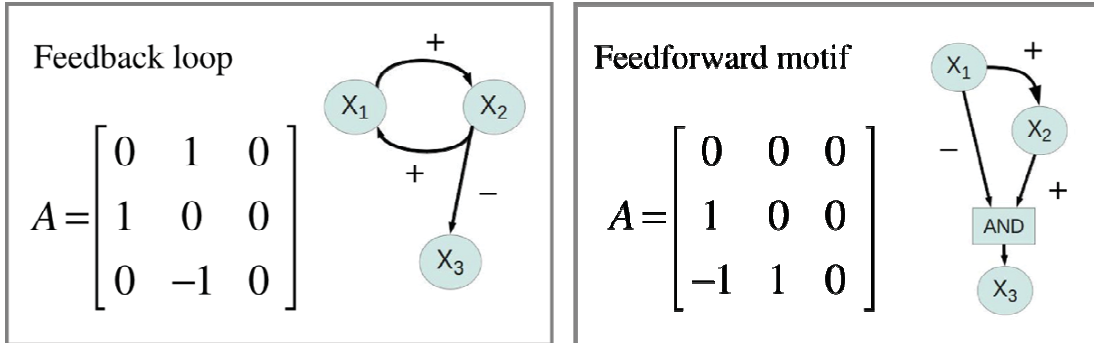


**Supplementary Figure 1 Measures of monotonicity versus additivity of GP maps with three loci.** Scatterplots showing  $V_A/V_G$  from unweighted regression versus (**left panel**) degree of monotonicity and (**right panel**) monotone  $R^2$ . Red dots show 1000 random three-locus GP maps, blue dots 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 (Gjuvslund et al, 2011).



**Supplementary Figure 2 Comparing measures of monotonicity GP maps.**

Scatterplots showing degree of monotonicity versus monotone  $R^2$ . 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 (Gjuvsland et al, 2011)).



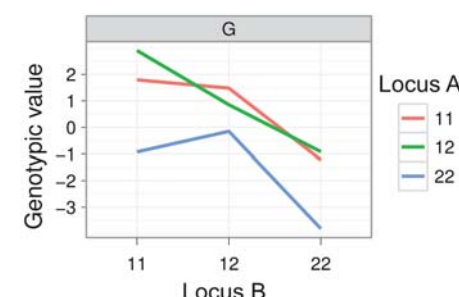
**Supplementary Figure 3 Connectivity matrices and signed directed graphs.** Two of the 1881 systems in the simulation study connectivity matrix  $A$  and the corresponding signed digraph of a system with a positive feedback loop between  $X_1$  and  $X_2$  while the right panel shows a system with incoherent feedforward from  $X_1$  to  $X_3$ .



## Supplementary Text 1 Measuring monotonicity of a GP map

In this supplementary text we provide details on measuring monotonicity of a genotype-phenotype (GP) map and code for computing this with the R package `gpmmap`. As an example GP map we use the data published by Cheverud & Routman (1995), for two loci underlying 10-week body-weight at 10 weeks in a mouse F<sub>2</sub> cross.

<b>Supplementary Table 1</b> (genotypic values from Table 1 in Cheverud & Routman 1995) Genotypic values for 10-wk body weight (in grams) at marker loci D7Mit17 (locus A) and DIMit7 (locus B) in a F <sub>2</sub> intercross of Large (LG/J; allele 2) and Small (SM/J; allele 1) inbred mouse strains			
	A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>
B <sub>1</sub> B <sub>1</sub>	36.839	37.951	34.118
B <sub>1</sub> B <sub>2</sub>	36.527	35.898	34.894
B <sub>2</sub> B <sub>2</sub>	33.824	34.125	31.234



**Supplementary Figure 4** Lineplot of the GP map in Supplementary Table 1.

In the main text we assumed for simplicity that the allele indexes is chosen such that G(1111) is the smallest of the homozygote genotypic values. This is not the case in this example, but since both monotonicity measures are invariant with respect to choice of indexes there is no need to reindex the alleles. Looking at the lineplot in Supplementary Figure 4 we find that this GP map is non-monotonic with respect to both loci, locus A shows marginal overdominance for the 11 and 22 genotypes at locus B, while locus B shows marginal overdominance for the 22 genotype at locus A. However the deviation from a monotone map is not extreme in the sense that there is a clear tendency that for both loci substituting allele 1 for allele 2 tends to increase the phenotype. Based on this qualitative assessment of the monotonicity of this GP map we expect  $m$  and  $R_{mono}^2$  values less than, but not much less than 1.

### Computing degree of monotonicity

We start with computing the set of single allele substitution effects for locus A by subtracting row 1 from row 2 and row 3 from in Supplementary Table 1 we get  $S^A = \{1.112, -3.833, -0.629, -1.004, 0.301, -2.891\}$  and divide this into sets of positive

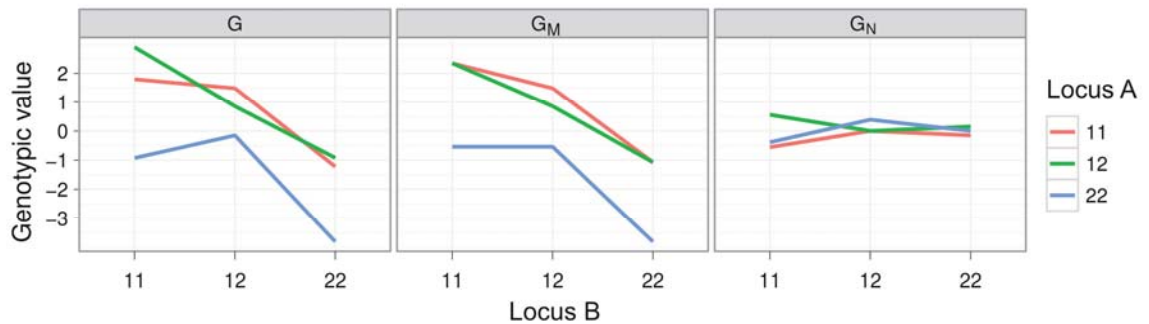
$S_+^A = \{1.112, 0.301\}$  and negative effects  $S_-^A = \{-3.833, -0.629, -1.004, -2.891\}$ . The sum of elements in  $S_+^A$  is 1.413 and the sum of absolute values of elements in  $S_-^A$  is 8.357 which gives  $P_A = 1.413$ ,  $N_A = 8.357$  and  $T_A = 9.770$ . From equation [5] in the main text we get  $m_A = 0.711$ . The sets of substitution effects for locus B can be found by subtracting rows in Table S1 to be  $S_-^B = \{-0.312, -2.703, -2.053, -1.773, -3.660\}$  and  $S_+^B = 0.776$ . This gives  $P_B = 0.776$ ,  $N_B = 10.501$ ,  $T_B = 11.277$  and  $m_B = 0.862$ . Inserting values for both loci into equation [6] in the main text we get the degree of monotonicity  $m = 0.792$ .

### Quantifying monotonicity by isotonic regression

Isotonic regression requires a partial ordering of genotype space as defined in equation [1] in the main text. In this two-locus example equation [1] defines 12 inequalities that can be laid out like this:

$$\begin{array}{ccccccc}
 1122 & < & 1222 & < & 2222 & & \\
 \mathbf{v} & & \mathbf{v} & & \mathbf{v} & & \\
 1112 & < & 1212 & < & 2212 & & \\
 \mathbf{v} & & \mathbf{v} & & \mathbf{v} & & \\
 1111 & < & 1211 & < & 2211 & & 
 \end{array} \tag{SI.1}$$

By isotone regression (Leeuw, Hornik et al. 2009) on the partial genotype ordering (SI.1) the original GP map is decomposed into a monotone and a non-monotone component (Supplementary Figure 54), with the coefficient of determination  $R_{mono}^2 = 0.97$ .



**Supplementary Figure 5** Lineplots of original GPmap G (as in Supplementary Figure 4) and its monotone ( $G_M$ ) and non-monotone ( $G_N$ ) components determined by isotone regression.

## R code

Reproducing numbers and figures in this text using the R package gpmap

```
> library(gpmap)           #load package
> data(GPmaps)             #load dataset
> gp <- mouseweight
>
> ## Table S1
> cbind(gp$genotype, gp$values)
>
> ## Supplementary Figure 4
> plot(gp)
>
> ## Computing degree of monotonicity
> gp <- degree_of_monotonicity(gp)
> gp$degree.monotonicity.locus
> print(gp)
>
> ## Quantifying monotonicity by isotonic regression
> gp <- decompose_monotone(gp)
> print(gp)
>
> ## Supplementary Figure 5
> plot(gp, decomposed=TRUE)
```

## References

Cheverud, J. M. and E. J. Routman (1995). "Epistasis and its contribution to genetic variance components." *Genetics* **139**(3): 1455-1461.

Leeuw, J. D., *et al.* (2009). "Isotone Optimization in R : Pool-Adjacent-Violators Algorithm ( PAVA ) and Active Set Methods." *Journal of Statistical Software* **32**(5).

# Supplementary Text 2

## Monotonicity is a key measure of genotype-phenotype maps

Arne B. Gjuvsland, Yunpeng Wang, Erik Plahte and Stig W. Omholt

In this supplement we first present a formal definition of the degree of monotonicity for a genotype-phenotype (GP) map. We then investigate monotonicity in the GP maps emerging from diploid gene regulatory networks. Under non-pleiotropic genotypic variation, i.e. genetic variation only in the maximum of the dose-response function, not in its shape, we show that the following is true:

1. If there are no positive feedback loops and no incoherent feedforward loops in the network, there is no order-breaking, the degree of monotonicity  $m = 1$ .
2. If there is a positive feedback loop or an incoherent feedforward loop in the network, there may be order-breaking in the GP map ( $m < 1$ ).

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.

### 1 Genotype-phenotype maps and degree of monotonicity

In this section we introduce a condensed notation for genotypes and genotype-phenotype maps and use it to give a more detailed presentation of the degree of monotonicity than in the main manuscript. To ease the presentation the present notation is slightly different in some respects.

We consider  $N$  diploid loci  $X_k$ ,  $k = 1, \dots, N$ , that together determine a real-valued phenotype. Each gene has two alleles with arbitrary indexes 1 and 2. We let  $g_k$  represent the genotype, i.e. the allelic composition of  $X_k$ . The two homozygous genotypes of  $X_k$  are  $g_k = 11$  and  $g_k = 22$ , the heterozygote is  $g_k = 12$ , in general  $g_k = \alpha_k \beta_k$ .

Multi-locus genotypes are made by concatenating single-locus genotypes. For instance, if an individual has  $g_j = 11$  and  $g_k = 12$ , then we denote the corresponding two-locus genotype by  $g_{jk} = g_j g_k = [1112]$ . The full  $N$ -locus genotype is  $g = [g_1 g_2 \dots g_k \dots g_N] = [g_{1:N}]$ . The space of all genotypes  $g$  is denoted  $\Gamma^N$ . For any locus  $X_k$ , the *genotypic background of  $X_k$* , i.e. the allele composition of all the genes *except*  $X_k$ , is  $g^{(k)} = [g_1 g_2 \dots g_{k-1} g_{k+1} \dots g_N] = [g_{1:k-1} g_{k+1:N}]$ . For example, if  $N = 4$  and  $j = 2$ , then  $g^{(2)} = [112212]$  means that the genotypes of  $X_1$ ,  $X_3$  and  $X_4$  are 11, 22, and 12, respectively. In the main paper we use (see

e.g. equation [1]) the straightforward notation  $[g_1g_2 \dots g_{k-1}11g_{k+1} \dots g_N] = [g_{1:k-1}11g_{k+1:N}]$  to indicate a genotype where  $g_j = 11$  while the background genotype is arbitrary. In the following we shall also use the compressed notation  $g = g_k g^{(k)}$ .

By a genotype-phenotype map we mean a mapping or transformation  $G : \Gamma^N \rightarrow R$  which maps a  $N$ -locus genotype  $g$  into a real-valued phenotype  $G(g)$ . For any given background  $g^{(k)}$  we define the two *allele substitution effects*  $s^1(g^{(k)})$  and  $s^2(g^{(k)})$  by

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

For example,  $s^1(g^{(k)})$  is the change in the genotypic value  $G(g)$  when allele 1 is substituted by allele 2 such that  $g_k$  changes from 11 to 12 while the background  $g^{(k)}$  is kept fixed.

We measure the monotonicity of  $G(g)$  with respect to locus  $X_k$  for a given background  $g^{(k)}$  by

$$\mu_k(g^{(k)}) = \frac{s^1(g^{(k)}) + s^2(g^{(k)})}{|s^1(g^{(k)})| + |s^2(g^{(k)})|}. \quad (2)$$

If both substitution effects are zero, we set  $\mu_k(g^{(k)}) = 1$ . Obviously,  $|\mu_k(g^{(k)})| \leq 1$ , the equality sign applying if and only if both differences have the same sign or if one or both of the terms are zero. Equivalently,  $|\mu_k(g^{(k)})| = 1$  iff  $G(g_{1:k-1}12g_{k+1:N}) \in [G(g_{1:k-1}11g_{k+1:N}), G(g_{1:k-1}22g_{k+1:N})]$ . Translated into single locus gene action terms this means that for a fixed background genotype  $g^{(k)}$ ,  $|\mu_k(g^{(k)})| = 1$  if locus  $X_k$  shows additivity, partial/complete dominance/recessivity or no phenotypic variation at all, whereas  $|\mu_k(g^{(k)})| < 1$  indicates overdominance or underdominance. Furthermore,  $\text{sign}(\mu_k(g^{(k)})) = \text{sign}(s^1(g^{(k)}) + s^2(g^{(k)})) = \text{sign}(G(g_{1:k-1}22g_{k+1:N}) - G(g_{1:k-1}11g_{k+1:N}))$ , i.e. the sign of  $\mu_k(g^{(k)})$  indicates which of the two homozygotes at locus  $X_k$  that give the higher phenotype for the given background.

The *degree of monotonicity* of  $G$  with respect to locus  $X_k$  is

$$m_k = \frac{\left| \sum_{g^{(k)}} \mu_k(g^{(k)}) t_k(g^{(k)}) \right|}{\sum_{g^{(k)}} t_k(g^{(k)})} = \frac{\left| \sum_{g^{(k)}} (s^1(g^{(k)}) + s^2(g^{(k)})) \right|}{\sum_{g^{(k)}} (|s^1(g^{(k)})| + |s^2(g^{(k)})|)}, \quad (3)$$

where  $t_k(g^{(k)}) = |s^1(g^{(k)})| + |s^2(g^{(k)})|$ , and each sum extends over all backgrounds  $g^{(k)}$ . The absolute value ensures that the measure  $m_k$  does not depend on our arbitrary numbering of the two alleles of  $X_k$ . 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 leave the value of  $m_k$  invariant.

By the triangle inequality  $m_k \leq 1$ . If  $m_k = 1$ , then  $G$  is monotonic with respect to locus  $X_k$ . On the other hand  $m_k < 1$  indicates non-monotonicity, which could occur for two different reasons. *Heterozygous non-monotonicity* occurs if there exists a background  $g^{(k)}$  for which  $|\mu_k(g^{(k)})| < 1$ , indicating that locus  $X_k$  shows over- or underdominance for the given background. *Homozygous non-monotonicity* occurs if the sign of  $\mu_k(g^{(k)})$  depends on the background  $g^{(k)}$ . This happens if there exist two backgrounds  $g_1^{(k)}$  and  $g_2^{(k)}$  such that  $G(22g_1^{(k)}) - G(11g_1^{(k)})$  and  $G(22g_2^{(k)}) - G(11g_2^{(k)})$  have opposite signs, i.e. the 11 homozygote gives the higher phenotype in one background while the 22 homozygote is the higher in the other, so-called *sign epistasis*. It follows that if  $G(g_{1:k-1}12g_{k+1:N}) \in [G(g_{1:k-1}11g_{k+1:N}), G(g_{1:k-1}22g_{k+1:N})]$  and  $G(g_{1:k-1}11g_{k+1:N}) \leq G(g_{1:k-1}22g_{k+1:N})$  for all  $g^{(k)}$  (or vice versa), then  $m_k = 1$ .

Finally, we define the overall degree of monotonicity  $m$  for the GP map  $G$  as a weighted average of all  $m_k$ . Denote the denominator in Eq. (3) by  $T_k$ . Then

$$m = \frac{\sum_{k=1}^N m_k T_k}{\sum_{k=1}^N T_k} = \frac{\left| \sum_{k,g^{(k)}} (s^1(g^{(k)}) + s^2(g^{(k)})) \right|}{\sum_{k,g^{(k)}} (|s^1(g^{(k)})| + |s^2(g^{(k)})|)}. \quad (4)$$

In the second expression the sums extend over all  $k$  and over all backgrounds  $g^{(k)}$  for each  $k$ . Obviously,  $m < 1$  if at least one  $m_k < 1$ , otherwise  $m = 1$ .

The present definition of  $m$  is equivalent to the one given in the main paper.

## 2 Order breaking, feedback loops and feedforward loops

In section 1 we studied monotonicity in generic genotype-phenotype maps. In the present section we analyse GP maps emerging from ODE models of gene regulatory networks. We consider a dynamic system consisting of  $N$  mutually interacting diploid loci  $X_i$ ,  $i = 1, \dots, N$ , regulating each other's expression. Following Omholt *et al.* (2000) we introduce genetic variation in regulatory parameters and use steady-state expression levels as phenotypes. In particular, we show that there are particular motifs in the network architecture whose presence tend to lower the degree of monotonicity by giving rise to order breaking.

We shall use the set notation for vectors and matrices introduced in Plahte *et al.* (tted). Let  $J$  and  $K$  be disjoint subsets of  $1, 2, \dots, N$ . Then for any vector  $v \in R^N$  and matrix  $A \in R^N \times R^N$ ,  $v_J$  is the vector with components  $v_j$ ,  $j \in J$ , and  $A_{JK}$  is the matrix with elements  $A_{jk}$ ,  $j \in J$  and  $k \in K$ . To indicate that some components or elements are excluded from a vector or a matrix, we use parentheses in superscripts. For example,  $x^{(j)} = [x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_N]$ , and  $A^{(jk)}$  is the matrix obtained from  $A$  by excluding row  $j$  and column  $k$ . In a similar way we let  $A^{(JK)}$  signify the matrix obtained from  $A$  by excluding the rows numbered by  $J$  and columns numbered by  $K$ .

The time dependent output of  $X_i$  is denoted  $z_i$ , and we define  $z = [z_1, z_2, \dots, z_N]$ . It goes without saying that  $z_i$  in general depends on the genotypes of all the genes even though we will not always state this explicitly. Our model, which was first suggested by Omholt *et al.* (2000), is based on the assumption that the total output rate of the gene is the sum of the output rates of each allele. This assumption is reasonable if the two alleles only differ in the regulatory domain of the gene, not in its coding domain.

For a given genotype  $g = \alpha_j \beta_j g^{(j)}$  the equations of motion for  $X_j$  are

$$\begin{aligned} \dot{z}_j^1 &= a_j^{\alpha_j} r_j^{\alpha_j}(z) - \gamma_j^{\alpha_j} z_j^1, \\ \dot{z}_j^2 &= a_j^{\beta_j} r_j^{\beta_j}(z) - \gamma_j^{\beta_j} z_j^2, \\ z_j &= z_j^1 + z_j^2, \end{aligned} \quad (5)$$

where  $z_j^{\alpha_j}$  and  $z_j^{\beta_j}$  are the time-dependent outputs of the two alleles. The two allele rate functions  $r_j^1(z)$  and  $r_j^2(z)$  have range  $[0, 1]$  so that  $a_j^1$  and  $a_j^2$  represent the maximum production rates of the two alleles. We assume that all dose-response functions in Eq. (5) 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.

In the following we are only concerned with the homostatic states of Eqs. (5), and assume for simplicity that they have just a single stable equilibrium point. Solving the stability conditions of Eq. (5) with respect to  $z_j^1$  and  $z_j^2$  and adding gives

$$f_j(x) = b_j^{\alpha_j} r_i^{\alpha_j}(x) + b_j^{\beta_j} r_j^{\beta_j}(x) - x_j = 0, \quad j = 1, \dots, N, \quad (6)$$

where  $x$  is the stable point,  $b_j^{\alpha_j} = a_j^{\alpha_j} / \gamma_j^{\alpha_j}$  and  $b_j^{\beta_j} = a_j^{\beta_j} / \gamma_j^{\beta_j}$ . Since our monotonicity measure is invariant with respect to the numbering of alleles, we will without loss of generality assume  $b_j^1 \leq b_j^2$  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 Eq. (6) by

$$J_{jk} = J_{jk}(g) = \frac{\partial f_j(x)}{\partial x_k}, \quad j, k = 1, \dots, N. \quad (7)$$

The determinant of  $J$  is  $D = \det(J)$ .

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  $X_i$  if and only if  $J_{ij} \neq 0$ , its sign given by the sign of  $J_{ij}$ . A chain or a directed path  $C_{ij}$  in  $\mathcal{G}$  is a set of arcs leading from  $X_j$  to  $X_i$ , all intermediate nodes being 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 (not to be confused with the signature of the loop, a quantity we shall not need here) is equal to the sign of  $L$ . Therefore, loops are either positive or negative. It is useful to distinguish between proper loops as defined above and composite loops. A composite loop is a union of two or more proper loops (the subloops of the composite loop) having no nodes in common. For example,  $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_1$  is a proper loop,  $(X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_1)(X_4 \rightarrow X_5 \rightarrow X_4)$  is a composite loop, while  $(X_1 \rightarrow X_2 \rightarrow X_1)(X_2 \rightarrow X_3 \rightarrow X_2)$  is not a loop at all, but two different loops involving  $X_2$ . The loop product of a composite loop is equal to the product of the loop products of its subloops. The feedback loops (proper and composite) in  $\mathcal{G}$  can be read out from  $J$  by computing its determinant and all its principal subdeterminants. Each nonzero term in any of these corresponds a feedback loop in  $\mathcal{G}$ . A node regulating itself constitutes an autoregulatory loop, corresponding to a nonzero diagonal element  $J_{ii}$  stemming from  $\partial f_i / \partial x_i \neq 0$  for some  $i$ . If  $\partial f_i / \partial x_i = -1$ , the rate function of node  $X_i$  is independent of  $x_i$ . In that case there is no effective autoregulation in a biological sense.

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).

In the following sections we analyse the causes of order breaking in the restricted case when there is only variation in the maximum transcription rates  $a_i^1$  and  $a_i^2$ , not in the shape of the dose-response functions  $r_i^1$  and  $r_i^2$ , implying  $r_i^1(x) = r_i^2(x) = r_i(x)$ . This restricted, *non-pleiotropic* kind of variation could occur by gene duplication or allele knockout in a homozygous locus.

As indicated in the first section, the system's phenotype could be any scalar quantity defined by its equilibrium value  $x$ . In the following we assume the genotype-phenotype map  $G(g) = x_q(g)$ ,  $q \in \{1 : N\}$ , for a given and fixed  $k$ , 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 next section we prove the following result:

**Proposition 1.** *Assume all rate functions in Eq. (5) are monotonic and that  $G = x_q$  for some fixed  $q$ . If there is no feedback loop (FBL) and no feedforward loop (FFL) anywhere in  $\mathcal{G}$  corresponding to the system Eq. (5), then necessarily  $m_k = 1$  for all  $k$ . In a system with a single feedback loop or a single feedforward loop there may be order breaking for some  $x_k$  if the feedback loop is positive or the feedforward loop 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  $x_q$ .

## 2.1 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  $X_i$ , there is no chain from  $X_i$  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  $X_l \rightarrow X_m$  with positive sign and no other chain from  $X_l \rightarrow X_m$ , 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

$$\begin{aligned} x_m(11g^{(j)}) &= 2b_m^1 r_m(x_l(11g^{(j)})), \\ x_m(12g^{(j)}) &= (b_m^1 + b_m^2) r_m(x_l(12g^{(j)})), \\ x_m(22g^{(j)}) &= 2b_m^2 r_m(x_l(22g^{(j)})). \end{aligned} \tag{8}$$

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

$$\begin{aligned} x_m(12g^{(j)}) &\geq (b_m^1 + b_m^2) r_m(x_l(11g^{(j)})), \\ x_m(22g^{(j)}) &\geq 2b_m^2 r_m(x_l(11g^{(j)})), \end{aligned} \tag{9}$$

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

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

$$\begin{aligned} x_j(11g^{(j)}) &= 2b_j^1, \\ x_j(12g^{(j)}) &= b_j^1 + b_j^2, \\ x_j(22g^{(j)}) &= 2b_j^2, \end{aligned} \tag{10}$$



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 the Lemma leads eventually to  $x_q(11g^{(j)}) \leq x_q(12g^{(j)}) \leq x_q(22g^{(j)})$ , irrespective of the genotypic background of  $X_j$ . 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, there can be no order breaking under non-pleiotropic genetic variation.

## 2.2 Networks with a feedback loop

In this section we investigate the effects of feedback loops on the degree of monotonicity. We consider a network in which there is no FFL and a single FBL. Without loss of generality let the loop be  $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_k \rightarrow X_1$ , comprising  $k$  nodes. There is at most one directed path from any node  $X_l$  to any other node  $X_m$ , and if there is a path from  $X_l$  to  $X_m$ , there is no return path from  $X_m$  to  $X_l$  if either  $X_l$  or  $X_m$  is not part of the loop.

In Plahte *et al.* (tted) we introduced *the propagation functions*  $x_k = p_{kj}(x_j)$  which express the effect on  $x_k$  of genetic variation in  $X_j$ . An important property of  $p_{kj}$  is that it can be derived from all the equilibrium conditions Eq. (6) except the equation for  $f_j$ . This implies that the effects of genotypic variation in  $X_j$  are only expressed in terms of  $x_j$ , while the parameters expressing the genotype of  $X_j$  do not enter into the function  $p_{kj}$ . We showed in the previous section that there can be no order breaking in a network without FBLs and FFLs. Below we show this again, using a more formal derivation, as a warm-up for the analysis of FBLs.

In a network without FBLs and FFLs let  $X_U$  be the set of nodes in the chain from  $X_k$  to  $X_k$ , initial node and terminal node included, and  $X_V$  the full set of nodes not in this chain. For example, if  $N = 5$  and the chain is  $X_2 \rightarrow X_1 \rightarrow X_4$ , then  $U = \{1, 2, 4\}$ ,  $V = \{3, 5\}$ ,  $X_U = \{X_1, X_2, X_4\}$ , and  $X_V = \{X_3, X_5\}$ . The derivative  $p'_{kk}(x_k)$  can be expressed as (Plahte *et al.*, ttet)

$$p'_{kk}(x_k) = (-1)^{n-1} \frac{D_{VV}}{D^{(jj)}} P, \quad (11)$$

where  $n$  is the number of nodes in the chain,  $D_{VV}$  is the principal subdeterminant of the Jacobian  $J$  of Eq. (6) composed of the rows and columns not in  $C$ ,  $D^{(jj)}$  is the principal subdeterminant of  $J$  with row  $j$  and column  $j$  deleted, and  $P$  is the product of the Jacobian elements corresponding to the arcs in the chain. As there is no feedback loop in any of the subsets of nodes defining these two subdeterminants, only the diagonal elements stemming from the term  $-x_i$  in  $f_i$  contribute, and  $D_{VV} = (-1)^{N-n}$ ,  $D^{(jj)} = (-1)^{N-1}$ , hence

$$p'_{kk}(x_k) = (-1)^{n-1+N-n+N-1} P = P. \quad (12)$$

The Jacobian element corresponding to the arc  $X_l \rightarrow X_m$  is  $J_{ml} = \partial r_m / \partial x_l$ , which by assumption has a fixed sign irrespective of the system's genotype. It follows that  $p'_{kk}$  has a fixed sign, and from this the conclusion that there can be no order breaking follows just as in the previous section.

Then assume there is a FBL in the network and first let  $X_q$  be a node in the loop. Without loss of generality let the loop be  $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_q \rightarrow X_1$ . Let  $X_k$  also be a member of the loop. We will show that if the loop is positive, the sign  $s_{qk}$  of  $\partial x_q / \partial b_k$  may vary with varying genotype  $g_k$ , while if the loop is negative,

$s_{qk}$  is fixed. In the following we suppress the dependence on  $x_i$  of any variable  $x_m$  whose value is unaffected by genetic variation in  $X_k$ .

As above,  $x_q = p_{qk}(x_k)$ , but also  $x_k = b_k r_k(x_{k-1})$  and  $x_{k-1} = p_{k-1,q}(x_q)$ . Combining these three equations leads to

$$x_q = p_{qk}(b_k r_k(x_{k-1})) = p_{qk}(b_k r_k(p_{k-1,q}(x_q))). \quad (13)$$

By assumption this equation is satisfied by the stable value  $x_q$ . For all  $i$ , let  $E_i$  be the equation Eq. (6). To derive the expressions for  $p_{qk}$  and  $p_{k-1,q}$  we use  $E_{k+1}, \dots, E_q$  and  $E_1, \dots, E_{k-1}$ , respectively. In other words,  $E_k$  is not used. It follows that the right-hand side of Eq. (13) does not depend implicitly on  $b_k$ . Using implicit differentiation on Eq. (13) therefore gives

$$\frac{\partial x_q}{\partial b_k} = \frac{\frac{dp_{qk}}{dx_k}}{1 - b_k \frac{dp_{qk}}{dx_k} \frac{dr_k}{dx_{k-1}} \frac{dp_{k-1,q}}{dx_q}}, \quad (14)$$

granted that the denominator in Eq. (14) is nonzero. Combining this with Eq. (12) leads to

$$\frac{\partial x_q}{\partial b_k} = \frac{P_C}{1 - L}, \quad (15)$$

where  $L$  is the loop product of the loop, i.e. the product of the Jacobian elements corresponding to all directed arcs in the loop, and  $P_C$  is the chain product for the chain from  $X_k$  to  $X_q$ .

It is easy to see that  $L$  is proportional to  $b_k$  and that  $b_k$  does not enter into  $P_C$ . If the loop is negative,  $L < 0$ , the sign of  $\partial x_q / \partial b_k$  is fixed and equal to the sign of  $P_C$ . If  $P_C > 0$ , then this implies that  $x_q(11g^{(k)}) \leq x_q(12g^{(k)}) \leq x_q(22g^{(k)})$  so that no order-break is possible, and likewise if  $P_C < 0$ . However, when the loop is positive, homozygous or heterozygous order breaking will occur if  $0 < L < 1$  for one background or one value of  $b_k$  and  $L > 1$  for another.

If  $X_k$  is not in the loop but upstream of some node  $X_j$  in the loop, genetic variation in  $X_k$  causes monotonicity in  $X_j$ , and the above analysis can be carried out more or less in the same way, also leading to Eq. (15). If  $X_q$  is not in the loop but downstream of some  $X_j$  in the loop, the order breaking in  $X_j$  may propagate to  $X_q$ , and monotonicity is not ensured.

### 2.3 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  $G = x_q$  of genetic variation in  $X_k$  by studying possible sign changes in  $s^1(g^{(k)})$  and  $s^2(g^{(k)})$  for varying background  $g^{(k)}$ . 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  $n_1$  and  $n_2$  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  $\{1, 2, \dots, N\}$ , and let  $V_1$  and  $V_2$  be their complements.

We again use the propagation function  $x_q = p_{qk}(x_k)$  which expresses the variation of  $x_q$  when  $x_k$  varies due to genotypic variation. Roughly speaking, the derivative of  $p_{qk}(x_k)$  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.*, tted). To the chain  $C_i$  we assign the chain weight  $w_i$  given by

$$w_i = (-1)^{n_i-1} \frac{D_{V_i V_i}}{D^{(kk)}}, \quad i = 1, 2, \quad (16)$$

where  $D_{V_i V_i}$  is the Jacobian subdeterminant for the nodes not included in  $C_i$ , and  $D^{(kk)}$  is the Jacobian subdeterminant for all nodes except  $X_k$ . Eq. (11) which expresses the the derivative of  $p_{qk}$  when there is a single chain from  $X_k$  to  $X_q$ , is then replaced by (Plahte *et al.*, tted)

$$\frac{dp_{qk}}{dx_k} = w_1 P_1 + w_2 P_2. \quad (17)$$

When there is no feedback loop in the system, only the diagonal elements in  $J$  stemming from the term  $-x_i$  in Eq. (6) contribute to the determinants  $D_{V_i V_i}$  and  $D^{(kk)}$ :

$$\begin{aligned} D_{V_i V_i} &= (-1)^{N-n_i}, \\ D^{(kk)} &= (-1)^{N-1}. \end{aligned} \quad (18)$$

Altogether this gives

$$\frac{dx_q}{dx_k} = \frac{dp_{qk}}{dx_k} = P_1 + P_2. \quad (19)$$

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  $dx_q/dx_k$  may vary. This may give rise to heterozygous as well as homozygous non-monotonicity. If the FFL is coherent, however, no order-breaking can occur.

If  $X_k$  is upstream relative to the initial node  $X_{\text{init}}$  of the FFL, it follows from Sect. 2.1 that there will be no order-breaking in  $X_{\text{init}}$ , and the above argument is still valid.

## 2.4 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  $\{1, 2, \dots, N\}$ , 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 Eq. (6) to

$$\begin{aligned} b_i^{\alpha_i} r_i^{\alpha_i}(x(g)) + b_i^{\beta_i} r_i^{\beta_i}(x(g)) - x_i(g) &= 0, \quad i = 1, \dots, N, \\ h(x_U(g)) - G(g) &= 0, \end{aligned} \quad (20)$$

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  $X_1, \dots, X_N$ . If the FFL is coherent, order breaking only occurs if it occurs in the original system.

## References

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**(2), 969–980.

Plahte, E., B., G. A., and Omholt, S. W. (submitted). Propagation of genetic variation in gene regulatory networks. *Submitted to Physica D: Nonlinear Phenomena*.

# **Paper III**



# Parameters in Dynamic Models of Complex Traits are Containers of Missing Heritability

Yunpeng Wang<sup>1</sup>, Arne B. Gjuvslund<sup>2</sup>, Jon Olav Vik<sup>2</sup>, Nicolas P. Smith<sup>3</sup>, Peter J. Hunter<sup>4</sup>, Stig W. Omholt<sup>1\*</sup>

**1** Centre for Integrative Genetics, Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway, **2** Centre for Integrative Genetics, Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås, Norway, **3** Department of Biomedical Engineering, St Thomas' Hospital, King's College London, London, United Kingdom, **4** Auckland Bioengineering Institute, The University of Auckland, Auckland, New Zealand

## Abstract

Polymorphisms identified in genome-wide association studies of human traits rarely explain more than a small proportion of the heritable variation, and improving this situation within the current paradigm appears daunting. Given a well-validated dynamic model of a complex physiological trait, a substantial part of the underlying genetic variation must manifest as variation in model parameters. These parameters are themselves phenotypic traits. By linking whole-cell phenotypic variation to genetic variation in a computational model of a single heart cell, incorporating genotype-to-parameter maps, we show that genome-wide association studies on parameters reveal much more genetic variation than when using higher-level cellular phenotypes. The results suggest that letting such studies be guided by computational physiology may facilitate a causal understanding of the genotype-to-phenotype map of complex traits, with strong implications for the development of phenomics technology.

**Citation:** Wang Y, Gjuvslund AB, Vik JO, Smith NP, Hunter PJ, et al. (2012) Parameters in Dynamic Models of Complex Traits are Containers of Missing Heritability. *PLoS Comput Biol* 8(4): e1002459. doi:10.1371/journal.pcbi.1002459

**Editor:** Frederick P. Roth, Harvard Medical School, United States of America

**Received:** September 12, 2011; **Accepted:** February 19, 2012; **Published:** April 5, 2012

**Copyright:** © 2012 Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the Research Council of Norway (<http://www.forskningradet.no/>) under the eVITA program, project number 178901/V30 and by the Virtual Physiological Rat Project funded through NIH grant P50-GM094503. NOTUR, the Norwegian metacenter for computational science, provided computing resources under project nn4653k. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [stig.omholt@umb.no](mailto:stig.omholt@umb.no)

## Introduction

The phenotypic variance cumulatively explained by marker loci found to associate with complex traits in genome-wide association studies (GWAS) is usually much less than the narrow-sense heritability [1–6], the ratio of additive genetic variance to total phenotypic variance. Several explanations have been proposed for this unexplained variance, popularly called the missing heritability [1], including imprecise heritability estimates; insufficient sample size; exclusion of particular types of polymorphisms such as copy number variants and rare SNPs in GWAS; unaccounted epistatic effects; and underestimated effect size of associated SNPs due to incomplete linkage with causal variants [3,6]. Recently it was shown [7] that a large proportion of the missing heritability can be accounted for if one estimates the variance explained by all available marker loci together. This suggests that most of the genetic variation underlying complex trait variation is due to marginal effects of many loci that are too small to pass stringent significance tests. Strong support for this interpretation comes from several meta-analyses of genome-wide association data [8–10]. While this insight appears to resolve much of the missing heritability issue as such, it also implies that standard GWAS approaches will not be very helpful for disclosing which genetic variants do actually contribute to additive variance.

Part of the problem underlying the missing heritability is that while the genotype-phenotype map in reality arises from complex biological systems best described by nonlinear dynamic models, the statistical machinery of quantitative genetics, including GWAS

methods, is built upon linear models of gene action. The aim of this study is not to improve the statistical methods *per se*, but rather to explore how more of the missing heritability can be explained and understood by combining nonlinear dynamic models with existing GWAS methods. The research program of linking system dynamics and genetics was suggested more than 40 years ago [11] and has been an active research area for more than 10 years [12–24]. Emergent properties of nonlinear systems, such as systemic silencing [25], might lead to a situation where genetic variation that penetrates to low-level phenotypes underlying a higher-level phenotype does not necessarily manifest in the higher-level phenotype itself. Doing GWAS on these low-phenotypes may thus reveal more of the genetic variation influencing the higher-level trait. This line of reasoning is reflected in recent GWA studies on metabolite profiles [26,27], in pathway and network-based analysis of genome-wide association studies [28] and in GWAS analyses on global gene expression data [29–30]. While all these studies represent important contributions, they do not combine a genetic mapping framework with mathematical models describing how high-level trait variation emerges from low-level trait variation, i.e. they do not provide a quantitative framework for elucidating how genetic variation affecting a low-level phenotype do actually influence a focal high-level phenotype.

If a dynamic model can describe the phenotypic variation of a given trait, it follows that irrespective of the biological resolution of the model, the genetic variation underlying the phenotypic variation will have to be reflected as variation in the parameters of the model. We therefore hypothesized that performing GWAS

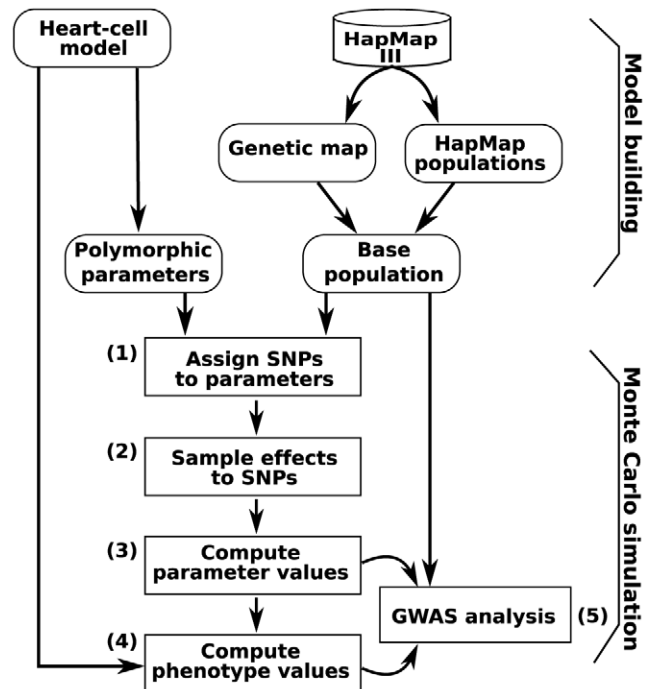
## Author Summary

Despite an ever-increasing number of genome locations reported to be associated with complex human diseases or quantitative traits, only a small proportion of phenotypic variations in a typical quantitative trait can be explained by the discovered variants. We argue that this problem can partly be resolved by combining the statistical methods of quantitative genetics with computational biology. We demonstrate this for the *in silico* genotype-to-phenotype map of a model heart cell in conjunction with publically accessible genomic data. We show that genome wide association studies (GWAS) on model parameters identify more causal variants and can build better prediction models for the higher-level phenotypes than by performing GWAS on the higher-level phenotypes themselves. Since model parameters are in principle measurable physiological phenotypes, our findings suggest that development of future phenotyping technologies could be guided by mathematical models of the biological systems being targeted.

on parameters in computational physiology models might reveal much more of the underlying genetic variation, as well as shedding light on how this variation actually causes phenotypic variation.

To test the plausibility of this reasoning, we combined GWAS methodology with a causally-cohesive genotype-phenotype (cGP) model linking genetic variation to phenotypic variation. More specifically, a cGP model [19] is a mathematical model of a biological system where low-level parameters have an articulated relationship to an individual's genotype, and higher-level phenotypes emerge from the mathematical model describing the causal dynamic relationships between these lower-level processes. Our approach bears some resemblance to that of functional GWAS (fGWAS) [31], where the genetic control of traits is analyzed by integrating biological principles of trait formation into the GWAS framework through mathematical and statistical bridges. Fu *et al.* [23] recently extended the functional mapping framework [15] to handle cyclic phenotypes such as circadian rhythms by combining differential equations with functional mapping of QTLs. However, there are clear differences between functional mapping and the cGP approach. In functional mapping the phenotypic measurements are currently done at the systems level, while lower-level parameters are estimated by combining curve-fitting with more classical QTL methods. In contrast, the cGP approach advocated here focuses on measuring lower-level parameters based on the idea that they are highly relevant phenotypes of the system.

We studied a cGP model of a mouse heart cell [24], where genetic variation is mapped to parametric variation, which propagates through the physiological model to generate multivariate phenotypes for the action potential (an electrical signal) and calcium transient (linked to muscle contraction) under regular pacing. The rationale for using a heart cell model was that multiscale and multiphysics modelling of the mammalian heart has a solid empirical basis, and arguably comprises the most complex mathematical conceptualization of any organ or physiological trait available. For clarity of exposition, and because heart cell models lie at the core of this class of multiscale whole organ models [32–37], we deemed it sufficient to illustrate our points using a single cell model only. We used HapMap data [38,39] as a guide to generate genetic variation with realistic allele frequencies and linkage disequilibrium to underlie variation in the model parameters. Based on HapMap [39] individuals we simulated complex pedigree populations and performed GWAS on both low-



**Figure 1. Flowchart of computational pipeline.** A heart cell model, a genetic map and a virtual population are tied together by selecting heart model parameters assumed to be under influence of genetic variation and associating the parameter variation to a population of virtual genomes based upon HapMap 3 data. Individual genotypes are mapped into heart model parameters (steps 1–3) and by running the heart cell model parameters are mapped into cell-level phenotypes (step 4). Finally, GWAS analysis is then performed on the virtual population (step 5).  
doi:10.1371/journal.pcbi.1002459.g001

level parameters and high-level phenotypes arising from the cGP model. The layout of the computational pipeline used for this study is depicted in Figure 1.

We show that genome-wide association studies on parameters reveal many more of the underlying SNPs than when using higher-level cellular phenotypes. Furthermore, the SNPs identified by GWAS on parameters can be used to build multivariate prediction models of higher-level phenotypes giving much higher explained variance than from GWAS on higher-level phenotypes alone. Our results suggest that combining statistical genetics with computational biology will facilitate both identification of genetic variation underlying complex traits and a much deeper understanding of how this genetic variation becomes causative.

## Methods

The general layout of the study is outlined in Figure 1.

### Heart cell model

The cell model [40] extends that of Bondarenko *et al.* [41] with more realistic calcium handling, conservation of charge, and detailed re-parameterization to consistent experimental data for the C57BL/6 “black 6” mouse. State variables include ion concentrations of sodium, potassium and calcium in the cytosol, calcium concentration in the sarcoplasmic reticulum, and the state distribution of ion channels, whose transition rates between open, closed, and inactivated conformations may depend on transmembrane voltage. Formulated as a system of coupled ordinary



differential equations, this model provides a comprehensive representation of membrane-bound channels and transporter functions as well as fluxes between the cytosol and intracellular organelles. As the action potential and calcium transient features following an electrical stimulation are the only state descriptors fed into higher level features of current multiscale heart models [33], we used these and associated aggregated measures as high-level phenotypes, see “Parameter to phenotype mapping” below. See Vik *et al.* [24] for a detailed description of this model including model diagram, differential equations and a CellML implementation.

### Polymorphic parameters

Out of the 86 model parameters we chose 34 to mediate the effects of genetic variation (Table 1 and Table S1). Because the genotype to parameter map for parameters describing ion channel properties may in general be much more straightforward than what is the case for many others, we picked mainly parameters describing affinities, conductivities and ion permeabilities for the ion channels and pumps underlying four potassium outward currents, one calcium current, one chloride current, one sodium current, the sodium-calcium exchangers, the sarcoplasmic reticular calcium ATPase (SERCA), the sodium potassium pump, cytosolic calmodulin, the ryanodine receptors on sarcoplasmic reticulum and the calcium handling processes within sarcoplasmic reticulum.

### Virtual genome and virtual population

To ensure some realism in the construction of the genetic structure of our *in silico* populations in terms of allele frequencies and LD patterns, we extracted HapMap3 data [39] for 2,000 evenly spaced SNPs (~5000 nucleotides apart) for each of the first 20 autosomal chromosomes. We extracted genotypes for the 40000 SNPs for the 1301 individuals in the 11 HapMap 3 populations. We then expanded this into a population of 5000 individuals by use of the Python package simuPOP [42]. The population expansion by simuPOP maintained allele frequencies and LD patterns in accordance with the HapMap 3 data. Mutations were introduced based on a symmetric diallelic mutation model, all recombinations were based on genetic maps estimated from the HapMap data and migrations between the 11 subpopulations were allowed for. The mutation rate, migration rate and number of generations used as input to the simuPOP population expansion were 1e-8, 0.001 and 500, respectively.

### Genotype to parameter mapping

Assuming a purely additive genetic model, 400 causal SNPs were randomly sampled from the virtual genome for each of the 34 parameters selected to mediate genetic variation. The genotype to parameter mapping for each parameter was set up by defining the 5,000×40,000 genotype matrix  $\mathbf{G}$ , where each element  $g_{ij}$  denoted the genotype of individual  $i$  at SNP  $j$  (−1 for the homozygous with the least frequent allele, 0 for the heterozygous and 1 for the homozygous with the most frequent allele). We then constructed for each parameter the 40,000×1 relative effect vector  $\mathbf{E}$ , where element  $e_j$  was sampled from a *Laplace* (0, 0.0035) distribution if the  $j$ -th SNP was among the 400 parameter-specific causative SNPs, and set to 0 otherwise (the relative effect being defined as the percentage increase or decrease of the baseline parameter value (Table 1 and Table S1)). The 5000-element vector of parameter values for all individuals was then computed as  $p(\mathbf{GE}+I)$ , where  $p$  is the baseline value. With this procedure, each of the focal 34 parameters was varied within  $\sim\pm 35\%$  of its baseline value, and for each causal SNP, the heterozygous

individuals were assigned the baseline parameter value (Table 1 and Table S1). The  $\pm 35\%$  parameter variation range was chosen as a compromise between getting ample genetic signals and avoiding too many physiologically unrealistic phenotypes. We also tested a genetic model with 200 causative SNPs for each parameter, the only difference being that the standard deviation of the *Laplace* distribution was set to 0.0049.

### Parameter to phenotype mapping

Cellular phenotypes for individual parameter sets were generated by a virtual experiment of constant pacing as described in Bondarenko *et al.* [41]. The potassium current was stimulated by −15 V/s for 3 ms at the start of each stimulus interval. Convergence was checked by comparing successive intervals with respect to the initial value of each state variable as well as the integral of its trajectory over that interval. A running history of 10 intervals was kept, and after each interval we checked for a match (within a relative tolerance of 5% for all state variables) against the previous one. This was done for three different pacing rates with stimulus intervals of intervals 100, 200 and 300 ms, respectively. The cell dynamics was categorized as “failure” if it did not converge to non-alternating dynamics within 10 minutes of simulation time. The Python code of the heart cell model was autogenerated from CellML [43], using the code generating service available at the CellML repository ([www.cellml.org](http://www.cellml.org)). The equations were integrated using the CVODE solver [44] with a Python wrapper.

Eight scalar phenotypes (see Table 2 and Table S2) were extracted from each computed action potential and calcium transient curve: the initial value (apbase and ctbase), the amplitude (apamp and ctamp), the peak value (apeak and ctpeak), the time to peak value (apttp and ctttp), the time to 25%, 50%, 75%, and 90% of the initial base value (apd25, apd50, apd75, apd90 and ctd25, ctd50, ctd75, ctd90).

### Data preparation

We removed individuals with physiologically unrealistic phenotypes within each of the 100 datasets analyzed. The exclusion criterion was based on the inter-quartile range (IQR); points that were more than twice the IQR above the third quartile or below the first quartile were excluded. Each filtered data set, containing 4000–5000 individuals, was divided into a training set of 2500 individuals and a test set consisting of the remaining individuals. The training data set was used to detect causal SNPs, compute the false positive rate and sensitivity characteristics. The test set was used to estimate the phenotypic variation accounted for by the detected SNPs.

### Statistical analysis

The same GWAS procedure was used for each parameter and each phenotype. The quantitative trait association analysis was performed with the program PLINK [45] on the training data. We used a threshold of 1e-5 on the Bonferroni-corrected p-value from PLINK to determine the set of significant SNPs.

The detected SNP set and associated discovery rates were defined as follows. Let  $\mathbf{S}_i$  denote the set of significant SNPs from GWAS on the  $i$ -th parameter and let  $\mathbf{C}_i$  denote the causal SNPs set of the  $i$ -th parameter. The set of detected SNPs of the  $i$ -th parameter was then computed as  $\mathbf{D}_i = \mathbf{S}_i \cap \mathbf{C}_i$ , and the discovery rate of  $i$ -th parameter was computed as  $d_i = |\mathbf{D}_i| / |\mathbf{C}_i|$ . The union of causal SNP sets for parameters defined the causal SNP set underlying all cellular phenotypes, and the detected SNP set and the discovery rate for each cellular phenotype was computed in the

**Table 1.** Parameters with genetic variation.

Parameter name	Description	Unit	Baseline value	Min	Max
<b>Ka+</b>	The PC1 – PO1 rate constant of the Ryanodine receptor	$\mu\text{M}^{-4}/\text{ms}$	6.08e-3	4.09e-3	8.06e-3
<b>Ka–</b>	The PO1 – PC1 rate constant of the Ryanodine receptor	$\text{ms}^{-1}$	7.133e-2	4.70e-2	9.58e-2
<b>Kb+</b>	The PO1 – PO2 rate constant of the Ryanodine receptor	$\mu\text{M}^{-3}/\text{ms}$	4.05e-3	2.64e-3	5.47e-3
<b>Kb–</b>	The PO2 – PO1 rate constant of the Ryanodine receptor	$\text{ms}^{-1}$	9.65e-1	6.32e-1	1.31
<b>Kc+</b>	The PO1 – PC2 rate constant of the Ryanodine receptor	$\text{ms}^{-1}$	9.00e-3	6.09e-3	1.20e-2
<b>Kc–</b>	The PC2 – PO1 rate constant of the Ryanodine receptor	$\text{ms}^{-1}$	8.00e-4	5.24e-4	1.07e-3
<b>m</b>	The $\text{Ca}^{2+}$ cooperativity parameter of PO1 – PO2 of the Ryanodine receptor	-	3.0	1.99	3.97
<b>n</b>	The $\text{Ca}^{2+}$ cooperativity parameter of PC1 – PO1 of the Ryanodine receptor	-	4.0	2.75	5.33
<b>P_CaL</b>	The permeability of the L-type $\text{Ca}^{2+}$ channel	$\text{ms}^{-1}$	2.5	1.62	3.30
<b>t_L</b>	The time constant of the switch between open and close states of the L-type $\text{Ca}^{2+}$ channel	$\text{ms}^{-1}$	1.5	1.01	1.98
<b>tau_L</b>	The inactivation time constant of the L-type $\text{Ca}^{2+}$ channel	$\text{ms}^{-1}$	1.15e3	7.82e2	1.52e3
<b>phi_L</b>	The proportion of closed states in open mode of the L-type $\text{Ca}^{2+}$ channel	-	1.80	1.23	2.43
<b>Kup</b>	The SERCA affinity to $\text{Ca}^{2+}$	$\mu\text{M}$	4.12e-1	2.93e-1	5.68e-1
<b>V1</b>	The leak constant of the Network Sarcoplasmic Reticulum	$\text{ms}^{-1}$	4.5	3.05	5.90
<b>KCSQN</b>	The Calsequestrin affinity to $\text{Ca}^{2+}$	$\mu\text{M}$	6.30e2	4.35e2	8.57e2
<b>K_Co</b>	The affinities of the $\text{Na}^+/\text{Ca}^{2+}$ exchanger to extracellular $\text{Ca}^{2+}$	$\mu\text{M}$	1.4e3	9.38e2	1.85e3
<b>K_Ci</b>	The affinities of the $\text{Na}^+/\text{Ca}^{2+}$ exchanger to intracellular $\text{Ca}^{2+}$	$\mu\text{M}$	3.6	2.45	4.93
<b>K_No</b>	The affinities of the $\text{Na}^+/\text{Ca}^{2+}$ exchanger to extracellular $\text{Na}^+$	$\mu\text{M}$	8.80e4	6.06e4	1.20e5
<b>K_Ni</b>	The affinities of the $\text{Na}^+/\text{Ca}^{2+}$ exchanger to intracellular $\text{Na}^+$	$\mu\text{M}$	1.2e4	8.38e3	1.58e4
<b>KNai</b>	The affinity of the $\text{Na}^+/\text{K}^+$ pump to intracellular $\text{Na}^+$	$\mu\text{M}$	1.66e4	1.13e4	2.17e4
<b>KKo</b>	The affinity of the $\text{Na}^+/\text{K}^+$ pump to extracellular $\text{K}^+$	$\mu\text{M}$	1.5e3	1.04e3	2.08e3
<b>KpCa</b>	The affinity of the $\text{Ca}^{2+}$ pump to intracellular $\text{Ca}^{2+}$	$\mu\text{M}$	2.89e-1	1.95e-1	3.93e-1
<b>Vmax</b>	The maximal exchange rate of $\text{Na}^+/\text{Ca}^{2+}$ exchanger	pA/pF	3.94	2.71	5.19
<b>Imax</b>	The maximal current of the $\text{Na}^+/\text{K}^+$ pump	pA/pF	2.49	1.71	3.58
<b>GK1</b>	The maximal conductance of the time-dependent $\text{K}^+$ channel	ms/ $\mu\text{F}$	3.5e-1	2.39e-1	4.52e-1
<b>GKr</b>	The maximal conductance of the rapid delayed rectifier $\text{K}^+$ channel	ms/ $\mu\text{F}$	1.65e-2	1.11e-2	2.17e-2
<b>GKur</b>	The maximal conductance of the ultrarapidly activating delayed rectifier $\text{K}^+$ channel	ms/ $\mu\text{F}$	2.50e-1	1.76e-1	3.27e-1
<b>KCl</b>	The half saturation constant of the $\text{Ca}^{2+}$ activated $\text{Cl}^-$ channel	$\mu\text{M}$	1.00e1	6.65	1.36e1
<b>GNa</b>	The maximal conductance of the $\text{Na}^+$ channel	ms/ $\mu\text{F}$	1.60e1	1.07e1	2.10e1
<b>GKtof</b>	The maximal conductance of the rapidly recovering transient outward $\text{K}^+$ channel	ms/ $\mu\text{F}$	5.35e-1	3.97e-1	7.11e-1
<b>GCICa</b>	The maximum conductance of the $\text{Ca}^{2+}$ activated $\text{Cl}^-$ channel	ms/ $\mu\text{F}$	1.00e1	6.56	1.33e1
<b>on_rate</b>	The autophosphorylation rate of Calmodulin	$\text{ms}^{-1}$	5.0e-2	3.25e-2	6.56e-2
<b>off_rate</b>	The dephosphorylation rate of the Calmodulin	$\text{ms}^{-1}$	2.0e-4	1.34e-4	2.67e-4
<b>IpCm</b>	The maximal current of the $\text{Ca}^{2+}$ pump	pA/pF	9.55e-2	6.35e-2	1.26e-1

Listing of the 34 parameters where genetic variation was introduced. The descriptions, units and baseline values are taken from the original publication [40]. The minimum and maximum values were obtained from the Monte Carlo simulations.  
doi:10.1371/journal.pcbi.1002459.t001

same way as for each parameter. The set of false positive SNPs of the  $i$ -th parameter or phenotype,  $\mathbf{F}_i$ , consists of SNPs in the set of significant SNPs  $\mathbf{S}_i$  that are not in the causal SNPs set  $\mathbf{C}_i$ . The false positive rate of the  $i$ -th parameter or phenotype was defined as the number of false positive SNPs in  $\mathbf{F}_i$  divided by the number of signals in  $\mathbf{S}_i$ ,  $|\mathbf{F}_i|/|\mathbf{S}_i|$ .

To quantify explained genetic variance a multiple regression model was constructed by regressing the phenotype or parameter value of the training set on the causal SNPs detected by GWAS (similar to the weighted genomic profile approach in [46]). Then the phenotypes of test set individuals were predicted using the corresponding fitted models. We measured the explained variation by the  $R^2$  values from regressing observed values on predicted phenotypic values for the individuals in the test set.

### Global sensitivity analysis

We quantified the linear sensitivity [47] of each phenotype to each parameter using linear regression in the training set. For each high-level phenotype and Monte Carlo simulation we used the 2500 simulated phenotypes as response and performed a series of univariate regressions each time with a single parameter as predictor. We measured global sensitivity by the coefficient of determination ( $R^2$ ).

### Results/Discussion

The proportion of true causative SNPs detected by GWAS was as expected substantially higher for the parameters than for the cellular phenotypes (Figure 2 and Figure S2 for the 200 SNPs

**Table 2.** Attained cellular phenotype values.

Phenotypes	Unit	Baseline value	Min	Max
apd25	ms	4.34	4.10	4.56
apd50	ms	5.89	5.33	6.39
apd75	ms	1.11e1	9.28	1.29e1
apd90	ms	1.95e1	1.60e1	2.30e1
apamp	mV	1.18e2	1.14e2	1.23e2
apbase	mV	-8.00e1	-8.06e1	-7.93e1
appeak	mV	3.82e1	3.41e1	4.23e1
apttp	ms	3.20	3.03	3.35
ctd25	ms	6.19e1	4.80e1	7.98e1
ctd50	ms	1.05e2	7.98e1	1.37e2
ctd75	ms	1.79e2	1.39e2	2.16e2
ctd90	ms	2.55e2	2.20e2	2.79e2
ctamp	μM	1.4e-1	4.85e-2	2.76e-1
ctbase	μM	8.14e-2	6.12e-2	1.05e-1
ctpeak	μM	0.22	1.15e-1	3.68e-1
ctttp	ms	2.40e1	1.93e1	2.98e1

The phenotypic values resulting from use of the baseline parameter values (see Table 1) are listed together with the minimum and maximum values achieved in the Monte Carlo simulations.  
doi:10.1371/journal.pcbi.1002459.t002

case). Median detection rates of causal SNPs were in the range 3.5%–4% after GWAS directly on parameter values (Figure 2A), and this number dropped to ~0.05% for GWAS studies on action potential phenotypes and ~0.02% for calcium transient phenotypes (Figure 2B), and the corresponding figures in the 200 SNPs case were 8–8.5%, ~0.16% and ~0.08%. The low overall detection rates were to be expected since we sampled SNP effects from an L-shaped distribution resulting in datasets where a small proportion of the SNPs underlying a given parameter will explain a substantial part of the variation. The main explanation for the decrease in detection rates is that the number of causal SNPs increases 34 times and the relative effects of all causal SNPs decrease, making them harder to pick up. Another, probably less important, phenomenon contributing to lower detection rates at the higher-level phenotypes is that going from parameter level to the system-level phenotype introduces nonlinearities in the SNP effects, and standard GWAS methods pick up only the additive part.

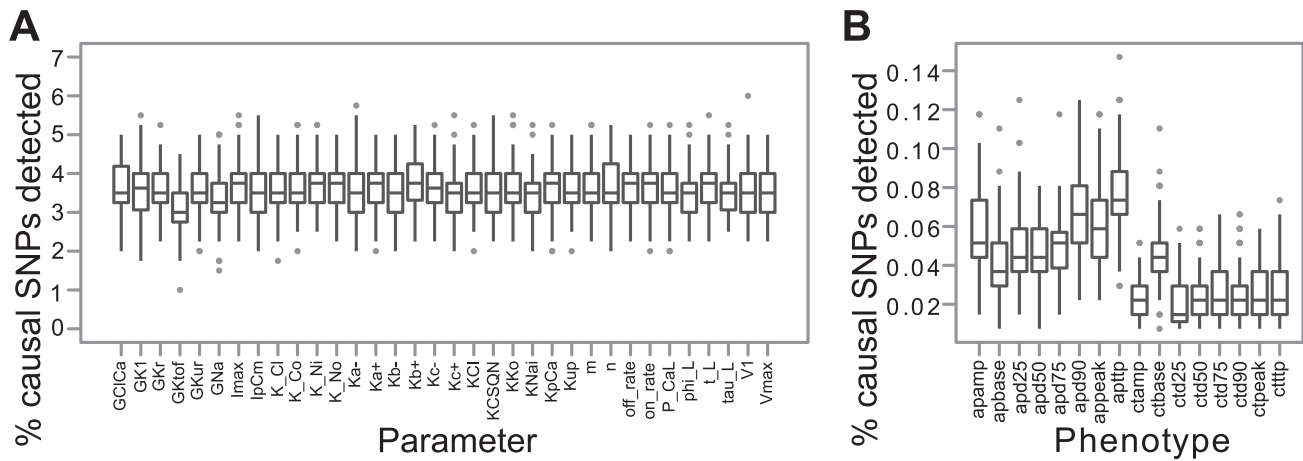
The difference between parameter and cellular phenotypes is also evident when looking at the amount of phenotypic variance explained by SNPs detected in the GWAS (Figure 3 and Figure S3 for the 200 SNPs case). The median explained variance is typically in the range 30–40% for parameter phenotypes (Figure 3A), 10–20% for action potential phenotypes and ~5% for calcium transient phenotypes (Figure 3B). The proportion of phenotypic variance explained by detected SNPs was on average 2.6 (2.0 in the 200 SNP case) and 5.6 (3.9 for the 200 SNPs case) times higher for a parameter phenotype than for an action potential and calcium transient phenotype, respectively. However, when we made use of the SNPs detected for parameters we were able to explain 1.8 and 3.9 times (1.6 and 2.9 times for the 200 SNPs case) more of the phenotypic variance of the action potential and calcium transient phenotypes, respectively, approaching the levels obtained for the parameters (Figure 3C). We also calculated the explained variances with all significant SNPs and obtained similar

results. This suggests that our approach can be tested empirically in a straightforward way.

The gain in explained variance by using parameter-associated SNPs was not as dramatic for the action potential phenotypes as for the calcium transient phenotypes (Figure 3C), but even in this case the gain in number of identified SNPs was on average  $13.9 \times$  ( $12.3$  for the 200 SNPs case). The corresponding figure for the calcium transient phenotypes was  $39.4 \times$  ( $26.5$  for the 200 SNPs case). Because these additional SNPs are attached to one or more parameters describing specific biological processes or features that are causally related according to the functional structure of the mathematical model, the gain in our causal understanding of the genotype to phenotype map may be substantial.

Both the detection rate of causal SNP and the variances explained for the calcium transient phenotypes were overall significantly lower than those for the action potential phenotypes (Figure 2B and 3B). We investigated this further by a linear global sensitivity analysis of how variation in the cellular phenotypes depended on variation in the parameters, and compared this with the number of causative SNPs for each parameter detected by performing GWAS on high-level cellular phenotypes. We found that the GWAS results for the two cellular phenotype groups are predominantly a consequence of the sensitivity structure of the dynamic model (Figure 4 and Figure S4 for the 200 SNPs case), and that the action potential phenotypes are overall more sensitive to fewer parameters than the calcium transient phenotypes. The only exception to this latter pattern is the parameter *Kup*, quantifying the affinity of SERCA to calcium ions (Figure 4A). It has a substantial impact on the calcium transient base value phenotype (*ctbase*), and the amount of variance explained by the SNPs detected for this phenotype is on par with the action potential phenotypes (Figure 3B). This suggests that SNPs associated with traits that are sensitive to few parameters will have a higher penetrance than SNPs associated with traits that are sensitive to many parameters for a given model resolution. Moreover, the results imply that the more poly-parametric the sensitivity profile of a model phenotype is, the more will be gained in terms of added explained variance by performing GWAS on parameters. On the other hand, the results also imply that a sensitivity analysis can be used to systematically reveal hotspots for genetic variation underlying a complex trait and thus guide a parameter phenotyping program. Within this framework a SNP affecting a parameter to which the focal higher-level phenotypes are not very sensitive will have little impact on these phenotypes unless it is highly penetrant at the parameter level.

GWAS methods are well known for producing false positives due to multiple testing and high LD between SNPs. A typical GWAS block of SNPs in high LD is often reduced to a subset of tagSNPs in low LD (typically with a pairwise correlation  $<0.2$ ). The GWAS methods are aimed at identifying significant tagSNPs, and the task of distinguishing the causal SNPs from false positives in high LD has to be done with other methods such as functional studies of candidate SNPs. Our approach is not intended to solve this problem (but see e.g. [48,49] for reviews of methods for identifying causal variants after GWAS) and in our study the increases detection rate for parameters is accompanied by an expected increased false positive rate (Figure S1 and Figure S5 for the 200 SNPs case). However, as parameters as a rule are closer to mechanism than higher-level phenotypes, it should be noted that to do GWAS on parameters could become very instrumental for identifying candidate mechanisms and genes for follow up studies. We envision that ongoing efforts such as the RICORDO project [50] aimed at developing semantic interoperability for biomedical data and models will facilitate bioinformatic identification of

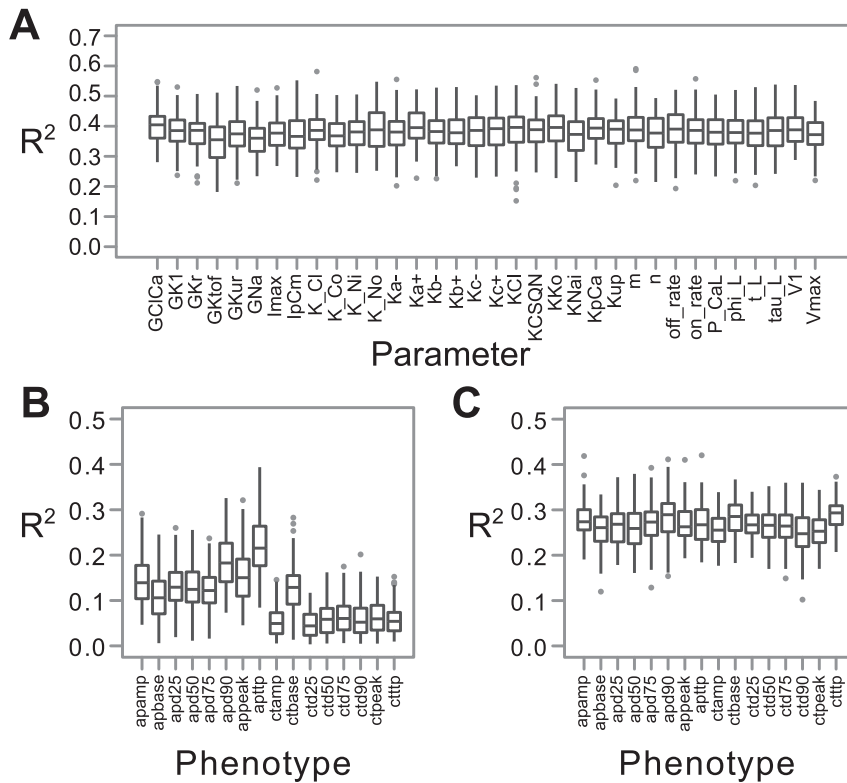


**Figure 2. Percentage of causative SNPs detected by GWAS.** (A) The percentage of 400 causative SNPs (y axis) detected as significant SNPs by GWAS on genetically controlled model parameters (x axis). (B) The percentage of all 13600 causative SNPs (y axis) detected as significant SNPs by GWAS on cellular phenotypes (x axis). Each boxplot summarizes 100 Monte Carlo runs. See Methods for further descriptions of model parameters and phenotypes.  
doi:10.1371/journal.pcbi.1002459.g002

candidate mechanisms and genes from cGP model sensitivities and GWAS results on parameter phenotypes.

We made deliberate use of the simplest possible genotype to parameter map in this study. A more complex map incorporating genetic dominance and various types of epistasis [51] would have

diminished the SNP discovery rates and the explained variances of the parameters. However, this reduction in penetrance would apply equally well at higher phenotypic levels, and so would not affect our conclusions. We did not put any environmental variation on the parameters as we deemed this unnecessary in a



**Figure 3. Phenotypic variance explained by genotypic variation.** (A) Total explained variance for genetically controlled parameters (x axis) using detected causal SNPs as predictors. (B) Total explained variance for cellular phenotypes (x axis) using detected causal SNPs obtained from GWAS targeting these phenotypes. (C) Total explained variance for cellular phenotypes (x axis) using detected causal SNPs obtained from GWAS targeting all genetically controlled parameters. Each boxplot summarizes total explained variance by GWAS for 100 Monte Carlo runs. Explained variance was measured as  $R^2$  from test set prediction with a multiple regression model, see Methods for further descriptions.  
doi:10.1371/journal.pcbi.1002459.g003



Table 2, the only difference being that it is based on 200 causal SNPs per parameter instead of 400.  
(PDF)

## References

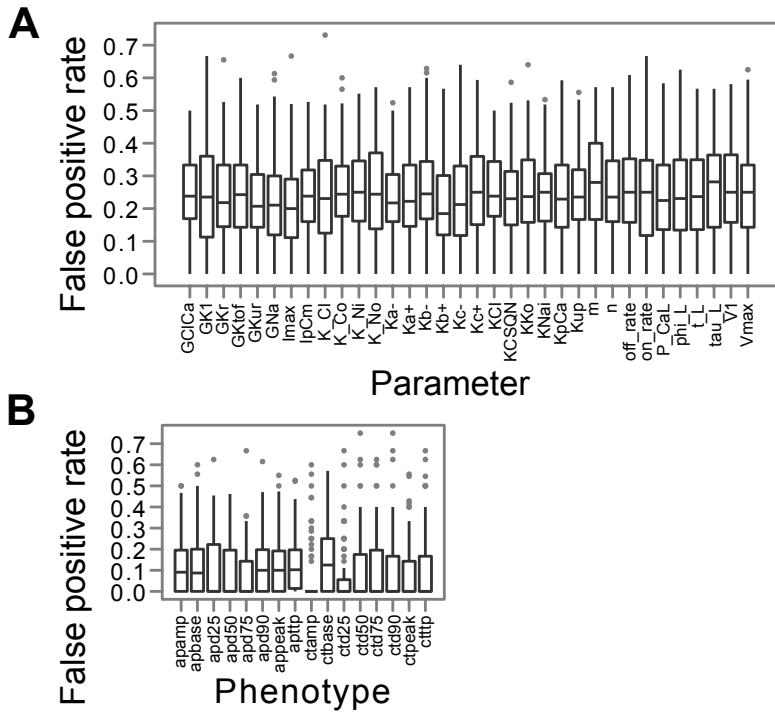
- Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456: 18–21. doi:10.1038/456018a.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461: 747–753. doi:10.1038/nature08494.
- Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. *PLoS Genet* 7: e1002051. doi:10.1371/journal.pgen.1002051.
- Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467: 832–838. doi:10.1038/nature09410.
- Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, 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.
- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11: 446–450. doi:10.1038/nrg2809.
- Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 43: 519–525. doi:10.1038/ng.823.
- Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40: 638–645. doi:10.1038/ng.120.
- Cho YS, Chen C-H, Hu C, Long J, Hee Ong RT, et al. (2011) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* 44: 67–72. doi:10.1038/ng.1019.
- Kato N, Takeuchi F, Tabara Y, Kelly TN, Go MJ, et al. (2011) Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat Genet* 43: 531–538. doi:10.1038/ng.834.
- Burns J (1970) The synthetic problem and the genotype-phenotype relation in cellular metabolism. In: Waddington CH, ed. *Towards a Theoretical Biology*. 3. Drafts. An I.U.B.S. Symposium. Chicago: Aldine Publishing Company. pp 47–51.
- Frank SA (1999) Population and Quantitative Genetics of Regulatory Networks. *J Theor Biol* 197: 281–294. doi:10.1006/jtbi.1998.0872.
- Omholt SW, Plahte E, Oyeaug L, Xiang K (2000) Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* 153: 969–980.
- Gilchrist MA, Nijhout HF (2001) Nonlinear developmental processes as sources of dominance. *Genetics* 159: 423–432.
- Ma C-X, Casella G, Wu R (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* 161: 1751–1762.
- Peccoud J, Velden KV, Podlich D, Winkler C, Arthur L, et al. (2004) The selective values of alleles in a molecular network model are context dependent. *Genetics* 166: 1715–1725.
- Welch SM, Dong Z, Roe JL, Das S (2005) Flowering time control: gene network modelling and the link to quantitative genetics: Modelling complex traits for plant improvement. *Aust J Agric Res* 56: 919–936.
- Gjuvsland AB, Hayes BJ, Omholt SW, Carlborg O (2007) Statistical epistasis is a generic feature of gene regulatory networks. *Genetics* 175: 411–420. doi:10.1534/genetics.106.058859.
- Rajasingh H, Gjuvsland AB, Våge DI, Omholt SW (2008) When parameters in dynamic models become phenotypes: a case study on flesh pigmentation in the chinook salmon (*Oncorhynchus tshawytscha*). *Genetics* 179: 1113–1118. doi:10.1534/genetics.108.087064.
- Gertz J, Gerke JP, Cohen BA (2010) Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions. *Theor Popul Biol* 77: 1–5. doi:10.1016/j.tpb.2009.10.002.
- Salazar-Ciudad I, Jernvall J (2010) A computational model of teeth and the developmental origins of morphological variation. *Nature* 464: 583–586. doi:10.1038/nature08838.
- Pumir A, Shraiman B (2011) Epistasis in a Model of Molecular Signal Transduction. *PLoS Comput Biol* 7: e1001134.
- Fu G, Wang Z, Li J, Wu R (2011) A mathematical framework for functional mapping of complex phenotypes using delay differential equations. *J Theor Biol* 289: 206–216. doi:10.1016/j.jtbi.2011.08.002.
- Vik JO, Gjuvsland AB, Li L, Tøndel K, Niederer S, et al. (2011) Genotype-phenotype map characteristics of an in silico heart cell. *Front Physiol* 2: 106. doi:10.3389/fphys.2011.00106.
- Gjuvsland AB, Plahte E, Omholt SW (2007) Threshold-dominated regulation hides genetic variation in gene expression networks. *BMC Syst Biol* 1: 57. doi:10.1186/1752-0509-1-57.
- Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, et al. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* 4: e1000282. doi:10.1371/journal.pgen.1000282.
- Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, et al. (2009) A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* 42: 137–141. doi:10.1038/ng.507.
- Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, et al. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* 18: 2078–2090. doi:10.1093/hmg/ddp120.
- Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10: 184–194. doi:10.1038/nrg2537.
- Ala-Korpela M, Kangas AJ, Inouye M (2011) Genome-wide association studies and systems biology: together at last. *Trends Genet* 27: 493–498. doi:10.1016/j.tig.2011.09.002.
- Das K, Li J, Wang Z, Tong C, Fu G, et al. (2011) A dynamic model for genome-wide association studies. *Hum Genet* 129: 629–639. doi:10.1007/s00439-011-0960-6.
- Noble D (2002) Modeling the Heart—from Genes to Cells to the Whole Organ. *Science* 295: 1678–1682. doi:10.1126/science.1069881.
- Smith NP, Mulquoney PJ, Nash MP, Bradley CP, Nickerson DP, et al. (2002) Mathematical modelling of the heart: cell to organ. *Chaos, Solitons & Fractals* 13: 1613–1621. doi:10.1016/S0960-0779(01)00170-9.
- Smith NP, Nickerson DP, Crampin EJ, Hunter PJ (2004) Multiscale computational modelling of the heart. *ANU* 13: 371–431. doi:10.1017/S0962492904000200.
- Hunter PJ, Borg TK (2003) Innovation: Integration from proteins to organs: the Physiome Project. *Nat Rev Mol Cell Biol* 4: 237–243. doi:10.1038/nrm1054.
- Nickerson D, Nash M, Nielsen P, Smith N, Hunter P (2006) Computational multiscale modeling in the IUPS Physiome Project: Modeling cardiac electromechanics. *IBM J Res & Dev* 50: 617–630. doi:10.1147/rd.506.0617.
- Smith N, de Vecchi A, McCormick M, Nordsletten D, Camara O, et al. (2011) euHeart: personalized and integrated cardiac care using patient-specific cardiovascular modelling. *Interface Focus* 1: 349–364.
- Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, et al. (2003) The International HapMap Project. *Nature* 426: 789–796. doi:10.1038/nature02168.
- Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58. doi:10.1038/nature09298.
- Li L, Niederer SA, Idigo W, Zhang YH, Swietach P, et al. (2010) A mathematical model of the murine ventricular myocyte: a data-driven biophysically based approach applied to mice overexpressing the canine NCX isoform. *Am J Physiol Heart Circ Physiol* 299: H1045–H1063. doi:10.1152/ajpheart.00219.2010.
- Bondarenko VE, Szigeti GP, Bett GCL, Kim S-J, Rasmuson RL (2004) Computer model of action potential of mouse ventricular myocytes. *Am J Physiol Heart Circ Physiol* 287: H1378–403. doi:10.1152/ajpheart.00185.2003.
- Peng B, Amos CI (2010) Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics* 11: 442. doi:10.1186/1471-2105-11-442.
- Lloyd CM, Halstead MDB, Nielsen PF (2004) CellML: its future, present and past. *Prog Biophys Mol Biol* 85: 433–450. doi:10.1016/j.pbiomolbio.2004.01.004.
- Cohen S, Hindmarsh C (1996) CVODE, a stiff/nonstiff ODE solver in C. *Computers in physics* 10: 138–143.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575. doi:10.1086/519795.
- Aulchenko YS, Struchalin MV, Belonogova NM, Axenovich TI, Weedon MN, et al. (2009) Predicting human height by Victorian and genomic methods. *Eur J Hum Genet* 17: 1070–1075. doi:10.1038/ejhg.2009.5.
- Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, et al. (2008) Global sensitivity analysis: the primer John Wiley & Sons. 305 p.
- Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet* 86: 6–22. doi:10.1016/j.ajhg.2009.11.017.
- Ioannidis JPA, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 10: 318–329. doi:10.1038/nrg2544.

## Author Contributions

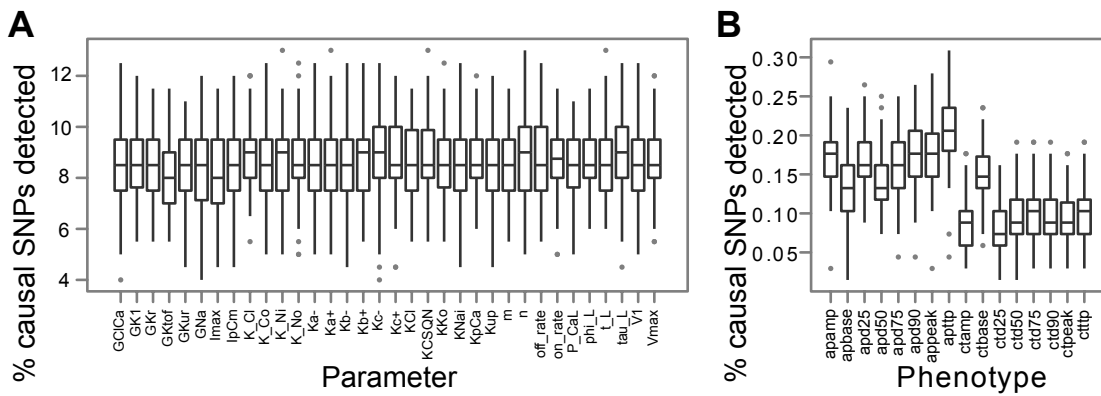
Conceived and designed the experiments: YW ABG SWO. Analyzed the data: YW ABG. Wrote the paper: YW ABG JOV NPS PJH SWO. Established computational pipeline: YW ABG JOV. Contributed to computational pipeline: NPS PJH.

50. de Bono B, Hochendorf R, Wimalaratne S, Gkoutos G, Grenon P (2011) The RICORDO approach to semantic interoperability for biomedical data and models: strategy, standards and solutions. *BMC Res Notes* 4: 313. doi:10.1186/1756-0500-4-313.
51. Phillips PC (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.
52. Dermitzakis ET, Clark AG (2009) Genetics. Life after GWA studies. *Science* 326: 239–240. doi:10.1126/science.1182009.
53. Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. *Nat Rev Genet* 11: 855–866. doi:10.1038/nrg2897.

**Figure S1**

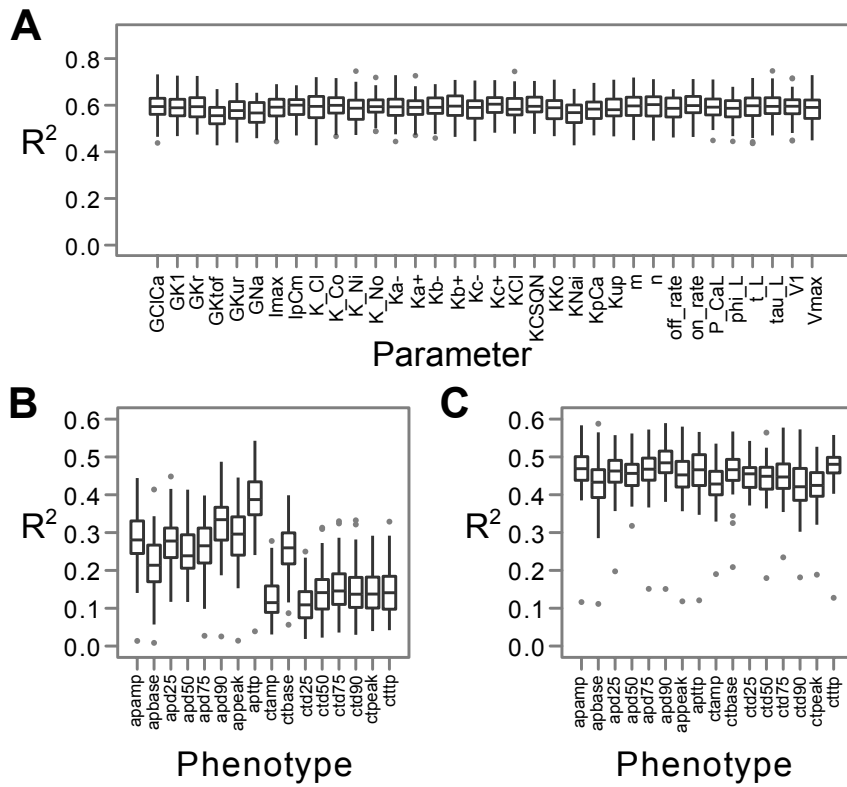


**Figure S2**





**Figure S3**



**Figure S4**

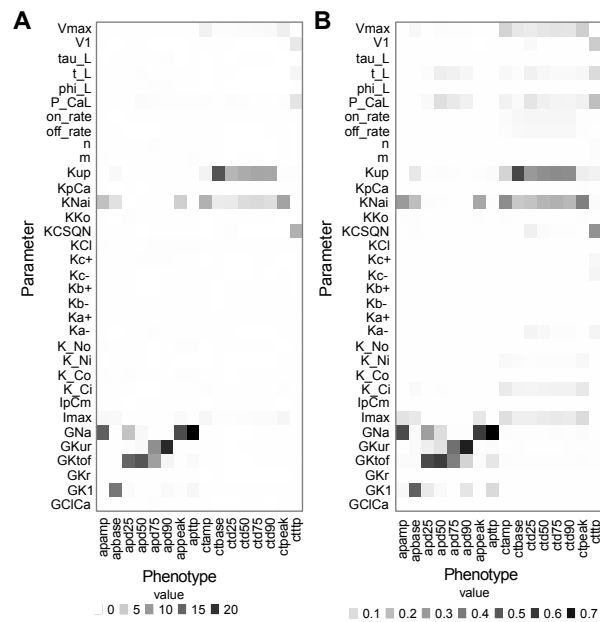


Figure S5

