Computational and Structural Biotechnology Journal 14 (2016) 28–34

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Uncovering the Genetic Architectures of Quantitative Traits

James J. Lee [a,*], Shashaank Vattikuti [b,*], Carson C. Chow [b,*]

[a] Department of Psychology, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA
[b] Mathematical Biology Section, NIDDK/LBM, National Institutes of Health, Bethesda, MD 20892, USA

ARTICLE INFO

Article history:
Received 10 August 2015
Received in revised form 16 October 2015
Accepted 23 October 2015
Available online 23 November 2015

Keywords:
Statistical genetics
Quantitative genetics
Population genetics
Average effect of gene substitution
Heritability
GWAS
Compressed sensing
Review

ABSTRACT

The aim of a genome-wide association study (GWAS) is to identify loci in the human genome affecting a phenotype of interest. This review summarizes some recent work on conceptual and methodological aspects of GWAS. The *average effect of gene substitution* at a given causal site in the genome is the key estimand in GWAS, and we argue for its fundamental importance. Implicit in the definition of average effect is a *linear model* relating genotype to phenotype. The fraction of the phenotypic variance ascribable to polymorphic sites with nonzero average effects in this linear model is called the *heritability*, and we describe methods for estimating this quantity from GWAS data. Finally, we show that the theory of *compressed sensing* can be used to provide a sharp estimate of the sample size required to identify essentially all sites contributing to the heritability of a given phenotype.

Contents

1. Introduction

The now-classic treatise *Genetics and the analysis of quantitative traits* [1], published three years before the first drafts of the human genome, covered the following sequence of topics:

1. definitions of key quantities in the study of quantitative (continuously varying) traits affected by multiple genetic and environmental causes,
2. methods for estimating some of these quantities without knowledge of the individual genetic sites affecting a given quantitative trait, and
3. the use of DNA-level data to identify the precise genomic regions that contain one or more such polymorphic sites.

In this review we survey work in all of these areas carried out in the decade and a half since the sequencing of the human genome. Modern genotyping technology has enabled genome-wide association studies (GWAS), which have led to a "golden age" of discovery in quantitative genetics [2], and we cannot hope to cover the substantial empirical progress in the identification of genetic loci contributing to quantitative variation. The most that can be done at the outset is to point the reader to the burgeoning research program in which our chosen conceptual and methodological issues are embedded [3–10].

Much of our discussion can be extended to binary phenotypes (such as disease diagnosis) through the device of treating liability as a quantitative trait affected by multiple genetic and environmental causes.

[*] Corresponding authors.
E-mail addresses: leex2293@umn.edu (J.J. Lee), vattikutis@niddk.nih.gov (S. Vattikuti), carsonc@niddk.nih.gov (C.C. Chow).

http://dx.doi.org/10.1016/j.csbj.2015.10.002

2. The Average Effect of Gene Substitution

We are interested in determining the quantitative influence of a polymorphic site on a given phenotype. Consider a biallelic site with alleles $\mathcal{A}_1$ and $\mathcal{A}_2$, where variation potentially affects a phenotype denoted by Y. A direct means to determine this quantity is to measure the phenotypic effect of experimentally changing the allelic state of the gene borne by a gamete. Confounding such an experiment, however, is dependence of the phenotypic effect on the allelic states of other genes in the zygote's genome. This nonlinear interaction is called *dominance* if it occurs between genes at the same site but inherited from different parents and *epistasis* if it occurs among genes at different sites. (We follow the classical usage of the term *gene* to refer to a token of heritable material at a given genomic site. Thus, each chromosome contains its own gene.) Fixing the allelic states everywhere else in the genome, we can write the effect of substituting $\mathcal{A}_2$ for $\mathcal{A}_1$, as

$$\Delta Y_{\mathcal{A}_1 \to \mathcal{A}_2 | \text{fixed background}} \tag{1}$$

It is not possible to estimate (1) for all backgrounds. There are roughly 10 million single-nucleotide polymorphisms (SNPs) in the human genome where the frequencies of both base pairs (alleles) exceed 0.01. Considering just these polymorphic sites alone, we have a number of multi-SNP genotypes equaling three to the power ten million. The developmental process maps each of these genotypes to an expected phenotypic value, but the astronomically large number of possible genotypes rules out any attempt to estimate this causal mapping in its totality. Even if a given genotype has a relatively high probability, in the sense of containing a common allele at each site, it is quite possible that no individuals in the population actually bear that genotype. Thus, even if it were possible to perform any conceivable mutagenic experiment [11], the sheer number of such experiments would place the genetic architecture of the phenotype—if this is defined by Eq. (1)—hopelessly out of our grasp.

We are thus forced to seek some more tractable object that preserves biological meaning. A natural thought is that we should concentrate on some weighted average of the possible gene substitutions at any given polymorphic site,

$$\alpha = \frac{\sum_k w_k \Delta Y_{\mathcal{A}_1 \to \mathcal{A}_2 | k}}{\sum_k w_k}, \tag{2}$$

where the sums are over all possible configurations (indexed by k) of alleles at the other genomic locations. The symbol α to represent the *average effect of gene substitution* was first used by Fisher [12]. The weights should take on the same values in the analogous expression defining the gene substitution $\mathcal{A}_2 \to \mathcal{A}_1$, such that these two quantities have the same absolute value but opposite signs.

Eq. (2) is an advance only if the weights allow the average to be calculated without knowledge of the myriad addends taking the form of Eq. (1). Fisher defined his average effect of gene substitution such that the weights reproduce the coefficient of the polymorphic site in the multiple regression of the phenotype on all such sites in the genome [13,14]. To make this equivalence more explicit, let **G** be the vector whose ith entry is the expected phenotype obtained by all organisms with a fixed multi-site genotype (arbitrarily labeled as the ith) developing within the current range of environmental conditions, **X** the matrix whose ijth entry is the number of genes (0, 1, or 2) of the jth allelic type present in the ith genotype, α the vector of average effects, and **R** the vector of residuals (Fig. 1). Without loss of generality, let all variables be standardized. Fisher effectively chose the weights in Eq. (2) such that the sum of the squared residuals,

$$\|\mathbf{R}\|_{\ell_2}^2 = \|\mathbf{G} - \mathbf{X}\alpha\|_{\ell_2}^2 \equiv \|\mathbf{G} - \mathbf{A}\|_{\ell_2}^2, \tag{3}$$

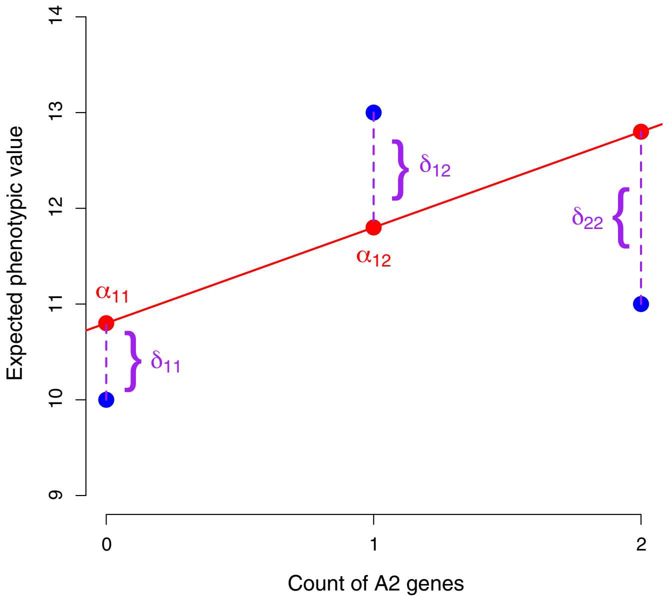

Fig. 1. Breeding (additive genetic) values and dominance deviations at a biallelic locus. The frequency of allele $\mathcal{A}_2$ is 0.6, and the causal effects of $\mathcal{A}_1\mathcal{A}_1 \to \mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_1\mathcal{A}_2 \to \mathcal{A}_2\mathcal{A}_2$ are 3 and −2 respectively. The genotype frequencies are in Hardy–Weinberg equilibrium. The phenotypic mean of each genotype is equal to the sum of its breeding value (α_{ij}) and genetic residual (δ_{ij}); in this case of nonlinearity within a locus, the genetic residuals are called *dominance deviations*. The phenotypic means are represented by the blue points, and the corresponding breeding values by the red points. The slope of the linear function giving the breeding values is the average effect of gene substitution.

is minimized. Eq. (3) defines a new quantity, $A_i = G_i - R_i = \sum_j X_{ij}\alpha_j$, the ith individual's so-called *breeding* or *additive genetic value*. The ℓ_2 norm is the *only* choice of norm in Eq. (3) that leads to the orthogonal decomposition of the total genetic variance,

$$\sigma_G^2 = \sigma_A^2 + \sigma_R^2. \tag{4}$$

All other choices will lead to the appearance of the covariance term 2 Cov(A, R), which essentially implies that the individual's breeding value does not contain all possible information about its phenotypic value that can be obtained from a linear combination of its single-site genotypes; some is abandoned in the residual. Thus, the choice of weights in Eq. (2) following from the use of the ℓ_2 norm in Eq. (3) is synonymous with the choice of variance as the measure of individual differences [15].

The variance in breeding value, σ_A^2, is called the *additive genetic variance*. The proportion of the total phenotypic variance, σ_Y^2, taken up by the additive genetic variance,

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}, \tag{5}$$

is called the *narrow-sense heritability* of the phenotype under consideration. When writers refer to “missing heritability,” they mean the discrepancy between estimates of Eq. (5) from studies of pedigrees and the percentage of the variance ascribable to phenotype-associated SNPs identified with high confidence in GWAS. Below, we will describe new methods for estimating h^2 and a means of identifying more of the SNPs contributing to this quantity.

In general, the weights in Eq. (2) are a difficult-to-compute function of the non-additive residuals, allele frequencies, and the correlation structure of polymorphic sites in the genome [14]. But it is of interest to examine the simplified case of a biallelic site that is uncorrelated—in *linkage equilibrium* (LE)—with all other causal sites and is itself in Hardy–Weinberg equilibrium. Let p_1 and p_2 denote the respective frequencies of $\mathcal{A}_1$ and $\mathcal{A}_2$. Suppose that we perform our hypothetical