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

COMPUTATIONAL AND STRUCTURAL 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 $\mathbf{G}$ 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, $\mathbf{X}$ 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 $\mathbf{R}$ 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,

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

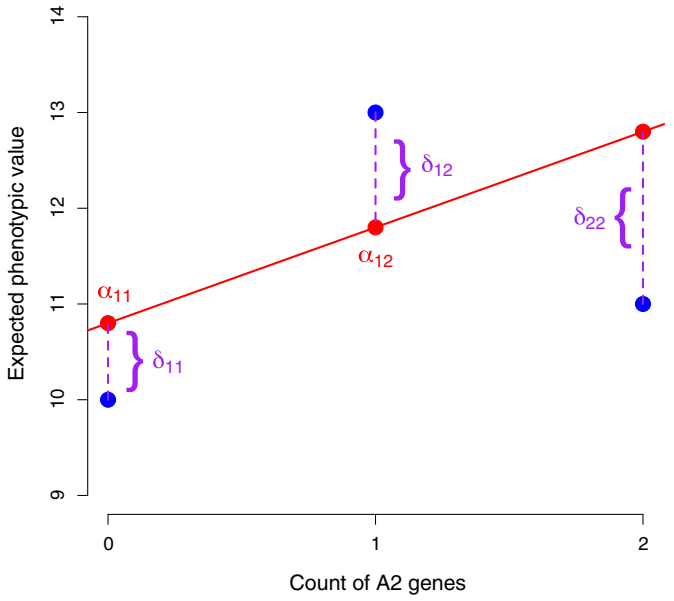

Fig. 1. Breeding (additive genetic) values and dominance deviations at a biallelic locus. The frequency of allele $\mathcal{A}_2$ is 0.6, and the causal effects of $\mathcal{A}_1\mathcal{A}_1 \to \mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_1\mathcal{A}_2 \to \mathcal{A}_2\mathcal{A}_2$ are 3 and − 2 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

mutagenic experiment on a randomly sampled gamete carrying a gene of the $\mathcal{A}_1$ allelic class. With probability p_1 its partner gamete will also carry $\mathcal{A}_1$, and with probability p_2 its partner gamete will carry the alternative $\mathcal{A}_2$. The expected effect of the gene substitution is thus

$$\frac{p_1 \Delta Y_{\mathcal{A}_1 \to \mathcal{A}_2 | \text{other gene is } \mathcal{A}_1} + p_2 \Delta Y_{\mathcal{A}_1 \to \mathcal{A}_2 | \text{other gene is } \mathcal{A}_2}}{p_1 + p_2}, \tag{6}$$

and it happens that in this case the weights (p_1, p_2) are precisely those leading to Fisher's average effect of gene substitution [16]. In reality it is likely that a causal site will be in linkage disequilibrium (LD) with other causal sites clustering near the same coding region. Distant causal sites may also be in very slight LD as a result of assortative mating or natural selection [14,15,17]. Nevertheless we think that the appealingly simple Eq. (6) will rarely give a poor approximation of the true average effect of gene substitution at a biallelic site.

3. The Linear Model of Quantitative Genetics

The concept of average effect is encapsulated in the linear model

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{R} + \mathbf{E}, \tag{7}$$

where $\mathbf{Y}$ is the vector of phenotypes, $\mathbf{X}$ is the genotype matrix, $\mathbf{R}$ is the vector of genetic residuals and $\mathbf{E}$ is the vector of non-genetic ("environmental") residuals.

We have tacitly assumed the absence of any correlation between the non-genetic residuals and any column of $\mathbf{X}$. Such confounding must be absent or remediable if we are to use empirical regression analysis to estimate the elements of α, as defined causally above. The inability to address analogous forms of confounding has been a bane to many fields of science limited to observational data [18]. A remarkable feature of GWAS, however, is that the correlation between the non-genetic residual and any given SNP is indeed often negligible [19]. We can point to a variety of checks supporting this claim, but perhaps the simplest and most convincing such check is the agreement between estimates of effects from samples of unrelated individuals and estimates from within families [5,8,20]. Recall that among the gametes produced by the same heterozygous parent, the allelic class of the transmitted allele is randomly selected and thus equivalent to treatment status in a randomized experiment [21,22]. A positive result in a within-family study thus provides powerful evidence that a SNP is indeed linked and associated with a site where the average effect is nonzero.

A potential objection to the linear model of quantitative genetics, which features coefficients that are averages over a large number of contexts, is that it sacrifices too much of biological interest for dubious gain. Holders of such a position tend to emphasize the importance of the full genetic architecture as represented by Eq. (1), although as a concession to the problem of combinatorial explosion they often begin with simplifying strategies such as limiting the first-pass analysis to pairwise interactions [23–25].

An important preliminary point is that scans for linear average effects (more or less standard GWAS practice) will not necessarily preclude the detection of causal sites that interact nonlinearly with each other. In order for a site involved in an epistatic interaction to exhibit an average effect equaling zero, the various terms in Eq. (2) must mutually cancel, which is an extremely unlikely occurrence.

The detection of sites with nonzero average effects thus serves as an excellent starting point even if the investigator's ultimate goal is the characterization of epistasis. There is an important respect, however, in which epistasis (defined in this quantitative–genetic sense) is less biologically significant than average effects. It turns out that nonlinear interactions do not make substantial contributions to familial resemblance.

Fig. 1 demonstrates this point in the case of a single causal site. The dominance deviations—nonlinear deviations of the conditional phenotypic means of the three genotypes from their corresponding breeding values—do not enter the correlations between ancestors and descendants [15]. To explain this remarkable fact, we start with the observation that dominance deviations are equivalent to the residuals in the least-squares linear regression of the conditional means on gene count. The residuals in any linear regression have an expected value of zero; the values of the outcome variable will show no systematic tendency to lie either above or below the regression line. If Hardy–Weinberg equilibrium holds, we can write this fact as

$$\sum_{i,j} p_i p_j \delta_{ij} = 0, \tag{8}$$

where δ_{ij} is the dominance deviation of the genotype with alleles $\mathcal{A}_i$ and $\mathcal{A}_j$ with respective probabilities p_i and p_j. Eq. (8) can be partitioned into terms that individually equal zero [26,27]. That is,

$$\sum_{j} p_i p_j \delta_{ij} = 0 \text{ for each } i, \tag{9}$$

which can also be put in the following way. In a subpopulation consisting of all individuals inheriting a particular allele (say $\mathcal{A}_1$) from a given parent (say the father), the mean of the dominance deviations is zero—just as in the population as a whole. The geometry of Fig. 1 should make this plausible. Since adjacent dominance deviations have opposite signs, the frequency-weighted sum of dominance deviations after fixing one allele will intuitively tend to cancel and in fact does so exactly.

Let us say that $\mathcal{A}_1$ is the allelic class of the gene that a parent transmits to its offspring. Under random mating the other gene at each individual's locus can be treated as drawn randomly from the entire population of genes. To simplify the notation, we now use p and $1 - p$ to denote the respective frequencies of $\mathcal{A}_1$ and $\mathcal{A}_2$. With probability $(1 - p)^2$, parent and offspring have the same dominance deviation δ_{11}. Similarly, with probability $2p(1 - p)$ they have different deviations (δ_{11} and δ_{12}), and with probability p^2 they share the heterozygous deviation (δ_{12}). Observe that

$$\begin{aligned} \mathrm{Cov}\left(\delta^{(\text{parent})}, \delta^{(\text{offspring})}\right) &= (1-p)^2\delta_{11}^2 + 2p(1-p)\delta_{11}\delta_{12} + p^2\delta_{12}^2 \\ &= \delta_{11}\left[(1-p)^2\delta_{11} + (1-p)p\delta_{12}\right] \\ &\quad + \delta_{12}\left[p(1-p)\delta_{11} + p^2\delta_{12}\right] \\ &= \delta_{11}\cdot 0 + \delta_{12}\cdot 0 \\ &= 0. \end{aligned} \tag{10}$$

It follows that the correlations between the phenotypes of ancestors and descendants are exactly the same regardless of whether the conditional phenotypic means of the possible genotypes actually lie on the line determined by the average effect or deviate nonlinearly.

This absence of nonlinear contributions to ancestor–descendant correlations does not generalize to all other forms of residual (non-additive) genetic variance. In particular, when there are interactions among genes at different loci, these can alter the correlations between relatives. However, these epistatic variance components have coefficients in the expression for a given correlation that decrease geometrically with the order of the interaction, and thus the great bulk of the contribution to the resemblance between relatives (other than monozygotic twins) continues to be made by the additive genetic variance. And this brings us to a commonsensical observation: if individual differences were caused primarily by non-additive genetic differences, then relatives would not strongly resemble each other, but it is unquestionably true that in our world relatives *do* resemble each other. This simple fact points to the importance and size of h^2, the proportion of the phenotypic variance due to variance in additive genetic value.

Given the undoubted importance of physical interactions between gene products in biological pathways, why do we not observe a more prominent role of epistasis in the genetic architectures of quantitative traits? One answer is that the typical allele frequencies at polymorphic sites may suppress the effects of the interactions that do occur. Once a new allele appears by mutation, the amount of time that it spends at each possible frequency p between zero and one before absorption at one of these two boundaries should be roughly proportional to $1/p$ [28], which means that we are much more likely now to observe the mutant when it is rare rather than common. This implies in turn that any genotype composed of many rare alleles must be much less common than its alternatives. One can appreciate the resulting tendency to linearize the genotype–phenotype mapping by inspecting Fig. 1. Suppose that the frequency of $\mathcal{A}_2$ evolves to be close to zero rather than 0.6. Then the homozygous genotype $\mathcal{A}_2\mathcal{A}_2$ will be so rare as to be given virtually no weight in the least-squares regression determining the average effect, and the regression line will then have to fit essentially only two points. An almost perfectly additive genetic architecture will have evolved out of an intrinsically nonlinear arrangement of the three conditional means. Likewise, in the case of multiple sites, the frequency spectrum of mutant alleles ensures that the least-squares hyperplane does not have to fit as many points as we might naively think [29,30]. Nonlinear architectures can be specially constructed to defeat this basic argument [24], but they require fine tuning [31].

Another answer is suggested by the striking concordance of GWAS findings across distinct populations. For instance, genetic effects from studies of East Asians are strongly correlated with estimates from studies of Europeans [32]. Because separately evolving populations differ in allele frequencies and LD patterns, the weights defining their respective average effects in Eq. (2) may be quite different. It seems to us that the simplest explanation for the agreement of the respective weighted averages despite the likely divergent weights is that the dependence on genomic background in Eq. (1) is often not very strong. This inference is explicable in light of a robust empirical regularity gleaned from GWAS: the individual effects of sites with common variants on a typical quantitative trait are quite small, often failing to account for even 1% of the phenotypic variance [2,33–35]. The heritability of a typical quantitative trait is thus spread across thousands of genomic sites, each accounting for a very small portion of $\mathrm{Var}(A)$. A fair conclusion to draw from this trend is that variation at a typical causal site perturbs the relevant biological system by a small amount. The smallness of individual effects implies even smaller nonlinear deviations from strict additivity [36].

4. Estimation of Heritability Using Unrelated Individuals

Having established that the average effect is the biologically relevant quantity to estimate, we now address how such quantities are estimated. The most straightforward approach is to estimate the average effects in Eq. (7) directly by regressing the phenotypes of a population against their genotypes. However, in real applications the number of imputed or sequenced polymorphic sites p will typically exceed the number of individuals in the dataset n. In so-called $p > n$ problems of this kind, the partial regression coefficients are not identifiable with ordinary least squares. In the next section, we show how the statistical theory of compressed sensing can be applied to directly estimate the individual average effects in the $p > n$ regime. Here, we show how an important aggregate quantity—h^2, the proportion of the phenotypic variance due to all genomic sites with nonzero average effects—can be estimated without knowledge of the individual sites contributing to this aggregate.

Classical methods of quantitative genetics estimate h^2 by determining the extent to which the correlations between relatives increases with the degree of biological relatedness. Under some simplifying assumptions the correlation between relatives is given by

$$\mathrm{Corr}\left(Y^{(\text{relative } i)}, Y^{(\text{relative } i')}\right) = A_{i,i'}h^2, \tag{11}$$

where $A_{i,i'}$ is a coefficient that depends on the pedigree relationship. For example, the coefficient equals unity if the relatives are monozygotic twins, 1/2 if they are parent and offspring, 1/4 if they are uncle (aunt) and nephew (niece), and so on.

The use of Eq. (11) to estimate h^2 from empirical correlations between relatives is often thought to be problematic because of the possibility that relatives resemble each other not only for genetic reasons but environmental ones [24]. This concern is probably overstated [37], but it is important to devise alternative estimators of h^2 so as to minimize the possibility that the so-called missing heritability is attributable to biases of pedigree studies.

Classical methods based on the correlations between relatives have been substantially augmented by a novel technique that makes use of GWAS data from nominally unrelated individuals [38,39]. This technique—often called *genomic-relatedness-matrix restricted maximum likelihood* (GREML) (we list URLs for all software tools at the end)—is perhaps the most important innovation in quantitative genetics to have been introduced in the last dozen years, and it has provided nearly definitive evidence for the view that undiscovered sites with common alleles account for a substantial portion of missing heritability.

For the moment we redefine the additive genetic variance, σ_A^2, to mean the variance that would be removed from the total phenotypic variance by multiple regression on all markers genotyped, sequenced, or imputed in a given study, as sample size goes to infinity. Because causal sites with a rare allele may not be present or represented by LD proxy in a given study, this additive genetic variance is less than the true additive genetic variance contributed by all polymorphic sites in the genome that we defined previously. Likewise, a site with a nonzero partial coefficient in the multiple regression now under consideration may not be a true causal site with a nonzero average effect but only an LD proxy for such a site. For convenience, however, we continue to use the terms "additive genetic variance," "heritability," "average effect" and their corresponding symbols in what follows.

We see from Eq. (7) that the total phenotypic variance can be written as

$$\begin{aligned}\mathrm{Var}(Y) &= \frac{1}{n}\mathrm{E}(\mathbf{Y}'\mathbf{Y}) \\ &= \frac{1}{n}\mathrm{E}(\alpha'\mathbf{X}'\mathbf{X}\alpha + \mathbf{e}'\mathbf{e}) \\ &= \sigma_A^2 + \sigma_E^2,\end{aligned} \tag{12}$$

where $\mathbf{e} = \mathbf{R} + \mathbf{E}$ and the expectation is over random $\mathbf{e}$. As before, the heritability is $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$. If we assume that LE holds approximately, then $\mathbf{X}'\mathbf{X} \approx n\mathbf{I}_p$ and the additive genetic variance is approximately $\alpha'\alpha$. We can see that Eq. (12) holds because $(1/n)E(\mathbf{u}'\mathbf{Z}'\mathbf{Z}\mathbf{u})$ is the variance of chip-based breeding values and hence equal to σ_A^2.

The goal is to estimate σ_A^2 given $\mathbf{X}$ and $\mathbf{Y}$. GREML treats Eq. (7) as the mixed-effects linear model

$$\begin{aligned}\mathrm{E}(\mathbf{Y}\mathbf{Y}') &= E(\mathbf{X}\alpha\alpha'\mathbf{X}' + \mathbf{e}\mathbf{e}') \\ &\simeq \mathbf{A}\sigma^2_{A,\mathrm{GREML}} + \mathbf{I}_n\sigma^2_{E,\mathrm{GREML}}\end{aligned} \tag{13}$$

and estimates the parameters $\sigma^2_{A,\mathrm{GREML}}$ and $\sigma^2_{E,\mathrm{GREML}}$, where, in the notation of [38], $\mathbf{A} = (1/p)\mathbf{X}\mathbf{X}'$ is the matrix of realized relatedness coefficients.

Eq. (13) is appealing because it assumes the same form as Eq. (11), except that the theoretical coefficient derived from the pedigree connecting biological relatives i and i' is replaced by the chance genetic similarity (which is either slightly greater or slightly less than a mean of zero) between essentially unrelated individuals [40]. Because the slight genetic similarities between unrelated individuals in a homogeneous population are not likely to be correlated with environmental similarities, it becomes safer to make the assumption above that breeding values are uncorrelated with the total residuals.

Despite the surface similarity between Eqs. (11) and (13), h^2 and h^2_{GREML} are not necessarily equal even under the same conditions that render Eq. (11) an unbiased estimator of h^2 [41]. The GREML Eq. (13) implicitly assumes that the outer product $\alpha\alpha'$ can be replaced by a diagonal matrix with all elements equal to the inner product $\alpha'\alpha$. As shown in [42] a sufficient condition for this approximation to be valid and as a result the equality of h^2 and h^2_{GREML} is that all sites are in LE. In practice, the two quantities will be very close if the causal sites are distributed randomly across the genome with respect to LD [42]. In other words, it must be the case that the extent of a site's LD with neighbors provides no information about its average effect (which may be zero). Since it is likely that causal variants tend to have lower minor allele frequencies (and hence are less well tagged by neighbors than a typical genotyped SNP) as a result of natural selection [33,35], we will usually have $h^2_{\text{GREML}} < h^2$. A number of methods have been proposed to bring these two quantities into close agreement regardless of minor allele frequency and LD [43–45]. It appears that the most robust means of addressing this issue is to form several different relatedness matrices, stratifying the SNPs by LD, and then to estimate the additive genetic variance as the sum of the scalars weighting the LD-defined relatedness matrices in the natural extension of Eq. (13) [46].

The GREML method and variants have been used to estimate the heritabilities of several human traits and also the genetic correlations between them. The *genetic correlation* is simply the correlation between the breeding values with respect to two phenotypes. [47] gives the model for estimation of the genetic correlation between two traits and [48] for the entire genetic correlation matrix of arbitrarily many traits. The multivariate applications of the GREML method have led to some of its most interesting results. For instance, it turns out that the genetic correlation between schizophrenia and bipolar disorder approaches 0.70 [49].

One advantage of GREML-type methods for heritability estimation over classical pedigree-based methods is that the former can partition heritability among different regions of the genome. Partitioning by chromosome has shown that the heritability contributed by each chromosome is often strongly correlated with its length [8,50], providing yet further evidence that the number of sites with nonzero average effects is typically very large. Partitioning by functional annotation has suggested that causal sites are disproportionately found in the vicinity of regions that are protein coding or DNase I hypersensitive [51]. Since the accuracy of the partitioning depends on the thoroughness of the imputation, these results should be taken as tentative. It is worth noting that both multivariate estimation and functional partitioning are more robust against LD than simple univariate estimation because of a tendency for biases to cancel from the numerators and denominators of the various estimands.

Very recently, a new method called *LD Score regression* has been introduced, and it can be put to some of the same uses as GREML [52–54]. When the chi-square statistics of the SNPs tested in a given GWAS are regressed against the "LD Scores" of the SNPs—the LD Score being a measure of the extent to which the focal SNP is in LD with its neighbors—the empirical result is an upwardly sloping straight line. This pattern is explicable in light of the fact that a SNP tagging more of its neighbors is thus more likely to tag one or more causal sites. Heuristically one might expect the value of the positive slope to provide an estimate of the trait's heritability, but the same GREML assumption regarding the absence of any relationship between average effect and LD must also hold for a valid estimate of h^2 to be obtainable from LD Score regression. (Others conditions may also be necessary.) For instance, if high-LD genomic regions tend to be devoid of causal SNPs, then the slope of LD Score regression will be biased downward (and the intercept biased upward).

In fact, the first use of LD Score regression suggested by its developers is not the estimation of heritability but rather the control of confounding. This use follows from the interpretation of the intercept as the expected chi-square statistic of a SNP with an LD Score of zero. The lowest possible LD Score of a SNP is in fact one, which is obtained when a SNP is in perfect LE with all other SNPs. This essentially means that a hypothetical SNP with an LD Score of zero fails to tag the average effect of any SNP in the genome, including whatever average effect the SNP itself may have. Therefore, if the intercept of LD Score regression departs upward from unity (the theoretical expectation of the chi-square distribution with one degree of freedom), the departure must be due to confounding, poor quality control, sample overlap, or other artifacts. This simple and ingenious method of estimating the distribution of truly null SNPs should in most cases lead to a much better global inflation of the association statistics than the overly conservative genomic control [55].

We close this section with some practical recommendations. In as-yet unpublished work, we have found that LD Score regression can return different heritability estimates than GREML even when applied to the same data. Thus, when the purpose is to estimate the heritability of a phenotype, GREML is the tool of choice since it is unbiased or can be made to be nearly so. In contrast, when the purpose is functional partitioning of heritability, we strongly recommend LD Score regression over GREML because the former method scales much better computationally with the number of categories to which the heritability is allocated. LD Score regression can also estimate a genetic correlation from the association Z-statistics of two traits, and here it also offers many advantages over GREML: computational speed, input consisting of summary statistics rather than individual-level data, and absorption of confounding into the intercept. So far LD Score regression has produced estimates of genetic correlations very similar to those yielded by GREML [54], and in our unpublished work it has also produced estimates very similar to those of an intuitive in-house method that is based on the simple correlation between the two vectors of marginal regression coefficients. As is the case with GREML, functional partitioning and bivariate estimation with LD Score regression are more robust than simple heritability estimation because of a tendency for biases to cancel from numerator and denominator.

5. Finding Trait-associated Genetic Markers With Compressed Sensing

For the vast majority of phenotypes studied so far, the majority of the sites with nonzero average effects contributing to the heritability have not yet been identified. We now discuss a particular means by which progress toward this goal might be advanced.

A typical GWAS evaluates millions of polymorphic sites (p). The number of subjects (n) is increasing dramatically, but $p > n$ will probably continue to hold for some time. As we stated earlier, the partial regression coefficients are not identifiable in this regime. Partly for this reason, GWAS investigators usually perform separate univariate regressions of their phenotype on each SNP and take forward the marginal coefficients obtained in this way. This approach is inherently unsatisfying, however, because the concepts of average effect and heritability rest on the partial coefficients. Therefore there is value in introducing some constraint (assumption) to deal with the ill-posed $p > n$ problem in the GWAS setting.

The Bayesian approach known as *genomic selection* (GS) depends on a prior distribution quantifying the assumption that most of the SNPs in a given panel have no average effect. A major drawback of this approach is the heavy computational cost of sampling methods for estimating the parameters of a Bayesian model. Reference [56] applied an approach based on combinatorial geometry and random matrix theory called *compressed sensing* (CS) [57–59], which, in contrast to the Bayesian approach, requires little more than the computationally tractable minimization of the lasso objective function

$$\|\hat{\mathbf{Y}}-\mathbf{Y}\|^2_{\ell_2} + \lambda\|\hat{\boldsymbol{\alpha}}\|_{\ell_1}, \tag{14}$$

where $\hat{\mathbf{Y}}$ is the estimated breeding value given by $\mathbf{X}\hat{\boldsymbol{\alpha}}$. The optimal choice of λ depends on the heritability contributed by the SNPs assayed in the

study, which can be estimated with GREML. The minimum of Eq. (14) over $\hat{\alpha}$ can be found efficiently with the pathwise coordinate optimization (PCO) algorithm [60]. In the case of LE, PCO has the same computational complexity as the standard GWAS approach, $O(np)$. LD increases the number of computations by either a constant or an amount that increases slowly with p (consistent with $\log p$). A memory-efficient implementation of lasso employing PCO is available in the latest version of PLINK [61].

Suppose that the number of nonzero elements in the true α is equal to s. CS theory shows that under fairly general conditions, if n is sufficiently large compared to s—but, crucially, not necessarily larger than p and perhaps much smaller—then the lasso or other ℓ_1-penalized schemes can select *all* polymorphic sites with nonzero coefficients in a multiple regression problem with high probability. (There is a major qualification, which we will explain shortly.) More specifically, if the sample size $n' < n$ is treated as a free parameter, then successive applications of the lasso to increasingly larger subsets of the data will result in a sharp transition from very poor selection to excellent selection. This transition can be observed in the behavior of the P-values returned by the standard univariate regressions of the phenotype on each of the SNPs selected by the lasso.

The CS approach makes no assumption about the distribution of the average effects. Instead it implicitly attempts to confine the estimate $\hat{\alpha}$ to an s-dimensional subspace. That is, if the true α in fact has $s \ll p$ nonzero elements, then these will be recovered by the lasso with high probability. There is evidence that, at least among sites where both alleles are common, $s \ll p$ for a wide range of traits [62,63]. Since n is expected to exceed s by a large factor even while falling well short of p, the prospects of recovering more heritability are quite promising, especially in light of the current push to generate large and widely available datasets. Note that although there is a relationship between ℓ_1-constrained solvers and the double Laplace prior that is debated in GS, CS theory is not based on this and holds for many different coefficient distributions and design matrices [57].

Finally, a given SNP is often strongly correlated—in tight LD—with several neighboring SNPs in the genome. This raises an obvious problem for the standard GWAS approach, since a causal SNP will lead many neighboring SNPs to exhibit nonzero univariate regression coefficients. The lasso does not in fact solve this problem. Although the lasso is statistically consistent under fairly general conditions, it may require a prohibitively large sample size to select only the causal sites in an LD block while setting the coefficients of all other sites to zero. Thus, in the presence of LD, "good recovery" means the selection of many sites that are false positives strictly speaking but nevertheless are in strong LD with one or more sites where the average effect is truly nonzero [56]. It is likely that no approach relying on statistical evidence alone can adequately address the problem of identifying the causal sites; external sources of biological evidence will be necessary. Particularly promising are empirical-Bayes approaches that use the trait-specific genome-wide relationship between GWAS signal and functional annotations (e.g., nonsynonymous status, tissue-specific DNase I hypersensitivity, chromatin modification, evolutionary conservation) to upweight the posterior probability of causality at certain sites [64,65].

6. Summary and Outlook

In this review we have argued that the average effect of gene substitution—a weighted average of the phenotypic changes that would result from idealized mutagenic experiments—is the pivotal quantity to be estimated in GWAS. Although this averaging may conceal important nonlinear effects of genetic variation on the focal phenotype, the identification of sites with nonzero average effects is at least an important starting point. In any event new methods of heritability estimation based on DNA-level data confirm classical findings from the correlations between relatives that much phenotypic variation is attributable to the average effects of gene substitution across all causal sites. Pinning down all of this additive genetic variance to individual locations in the genome with high confidence continues to be a challenge, since the average effects are typically very small, but the theory of CS provides reason to believe that a transition to good recovery is attainable with a combination of ℓ_1-penalization and large but reasonably realistic sample sizes.

Lurking not so far in the background behind all of these issues are the complications introduced by LD. Even if an oracle reveals to us the identity of a true causal site, that site's univariate regression coefficient may fail to equal its average effect of gene substitution as a result of LD. Perhaps a far more important concern is that LD prevents easy identification of causal sites responsible for GWAS signals in the first place. Furthermore, LD raises problems for GREML-type methods of heritability estimation that can probably stand further scrutiny. Notwithstanding these issues, however, the remarkable progress in quantitative genetics over the last decade leaves little doubt about the bountifulness of this research frontier.

URLs

GCTA-GREML, http://cnsgenomics.org/software/gcta;
LD Score regression, http://www.github.com/bulik/ldsc;
PLINK, https://www.cog-genomics.org/plink2.

References

[1] Lynch M, Walsh B. Genetics and the analysis of quantitative traits. Sunderland, MA: Sinauer; 1998.
[2] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet 2012;90(1):7–24. http://dx.doi.org/10.1016/j.ajhg.2011.11.029.
[3] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 2007;447(7145): 661–78. http://dx.doi.org/10.1038/nature05911.
[4] Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 2013;340(6139):1467–71. http://dx.doi.org/10.1126/science.1235488.
[5] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature 2014;511(7510): 421–7. http://dx.doi.org/10.1038/nature13595.
[6] Rietveld CA, Esko T, Davies G, Pers TH, Turley P, Benyamin B, et al. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. Proc Natl Acad Sci U S A 2014;111(38):13790–4. http://dx.doi.org/10.1073/pnas.1404623111.
[7] Perry JRB, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature 2014;514(7520):92–7. http://dx.doi.org/10.1038/nature13545.
[8] Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 2014;46(11):1173–86. http://dx.doi.org/10.1038/ng.3097.
[9] Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira B, Locke AE, Mägi B, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature 2015; 518(7538):187–96. http://dx.doi.org/10.1038/nature14132.
[10] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 2015;518(7538): 197–206. http://dx.doi.org/10.1038/nature14177.
[11] Pál C, Papp B, Pósfai G. The dawn of evolutionary genome engineering. Nat Rev Genet 2014;15(7):504–12. http://dx.doi.org/10.1038/nrg3746.
[12] Fisher RA. The genetical theory of natural selection. Oxford, UK: Oxford University Press; 1930.
[13] Fisher RA. Average excess and average effect of a gene substitution. Ann Eugen 1941; 11:53–63.
[14] Lee JJ, Chow CC. The causal meaning of Fisher's average effect. Genet Res 2013; 95(2–3):89–109. http://dx.doi.org/10.1017/S0016672313000074.
[15] Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 1918;52:399–433.
[16] Falconer DS. A note on Fisher's 'average effect' and 'average excess'. Genet Res 1985; 46(3):337–47. http://dx.doi.org/10.1017/S0016672300022825.
[17] Bulmer MG. The effect of selection on genetic variability. Am Nat 1971;105(943): 201–11.
[18] Freedman D. Statistical models and causal inference: a dialogue with the social sciences. New York, NY: Cambridge University Press; 2010.
[19] Lee JJ. Correlation and causation in the study of personality (with discussion). Eur J Personal 2012;26(4):372–412. http://dx.doi.org/10.1002/per.1863.
[20] Rietveld CA, Conley D, Eriksson N, Esko T, Medland SE, Vinkhuyzen B, et al. Replicability and robustness of genome-wide-association studies for behavioral traits. Psychol Sci 2014;25(11):1975–86. http://dx.doi.org/10.1177/0956797614545132.

[21] Fisher RA. Statistical methods in genetics. Heredity 1952;6:1–12.
[22] Ewens WJ, Li M, Spielman RS. A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. PLoS Genet 2008;4(9), e1000180. http://dx.doi.org/10.1371/journal.pgen.1000180.
[23] Hemani G, Theocharidis A, Wei W, Haley C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 2011;27(11): 1462–5. http://dx.doi.org/10.1093/bioinformatics/btr172.
[24] Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 2012;109(4): 1193–8. http://dx.doi.org/10.1073/pnas.1119675109.
[25] Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. PLoS Genet 2012;8(4), e1002625. http://dx.doi.org/10.1371/journal.pgen.1002625.
[26] Kimura M. On the change of population fitness by natural selection. Heredity 1958; 12(2):145–67. http://dx.doi.org/10.1038/hdy.1958.21.
[27] Moran PAP, Smith CAB. Commentary on R. A. Fisher's paper on the correlation between relatives on the supposition of Mendelian inheritance. London, UK: Cambridge University Press; 1966.
[28] Fisher RA. The distribution of gene ratios for rare mutations. Proc Roy Soc Edinb 1930;50:205–20.
[29] Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 2008;4(2), e1000008. http://dx.doi.org/10.1371/journal.pgen.1000008.
[30] Maki-Tanila A, Hill WG. Influence of gene interaction on complex trait variation with multilocus models. Genetics 2014;198(1):355–67. http://dx.doi.org/10.1534/genetics.114.165282.
[31] Stringer S, Derks EM, Kahn RS, Hill WG, Wray NR. Assumptions and properties of limiting pathway models for analysis of epistasis in complex traits. PLoS One 2013;8(7), e68913. http://dx.doi.org/10.1371/journal.pone.0068913.
[32] Marigorta UM, Navarro A. High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet 2013;9(6), e1003566. http://dx.doi.org/10.1371/journal.pgen.1003566.
[33] Chabris CF, Lee JJ, Benjamin DJ, Beauchamp B, Glaeser EL, Borst B, et al. Why it is hard to find genes that are associated with social science traits: theoretical and empirical considerations. Am J Public Health 2013;103(S1):S152–66. http://dx.doi.org/10.2105/AJPH.2013.301327.
[34] Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat Neurosci 2014;17(6):782–90. http://dx.doi.org/10.1038/nn.3708.
[35] Chabris CF, Lee JJ, Cesarini D, Benjamin DJ, Laibson DI. The fourth law of behavior genetics. Curr Dir Psychol Sci 2015;24(4):304–12. http://dx.doi.org/10.1177/0963721415580430.
[36] Crow JF. On epistasis: why it is unimportant in polygenic directional selection. Philos Trans R Soc B 2010;365(1544):1241–4. http://dx.doi.org/10.1098/rstb.2009.0275.
[37] Polderman TJC, Benyamin B, de Leeuw CA, Sullivan B, van Bochoven A, Visscher PM, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet 2015;47(7):702–9. http://dx.doi.org/10.1038/ng.3285.
[38] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010;42(7):565–9. http://dx.doi.org/10.1038/ng.608.
[39] Lee SH, Wray NR, Goddard ME, Visscher B. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88(3):294–305. http://dx.doi.org/10.1016/j.ajhg.2011.02.002.
[40] Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 2010;11(11):800–5. http://dx.doi.org/10.1038/nrg2865.
[41] Weir BS, Cockerham CC, Reynolds J. The effects of linkage and linkage disequilibrium on the covariances of noninbred relatives. Heredity 1980;45(3):351–9. http://dx.doi.org/10.1038/hdy.1980.77.
[42] Lee JJ, Chow CC. Conditions for the validity of SNP-based heritability estimation. Hum Genet 2014;133(8):1011–22. http://dx.doi.org/10.1007/s00439-014-1441-5.
[43] Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 2012;91(6):1011–21. http://dx.doi.org/10.1016/j.ajhg.2012.10.010.
[44] Lee SH, Yang J, Chen GB, Ripke S, Stahl EA, Hultman CM, et al. Estimation of SNP heritability from dense genotype data. Am J Hum Genet 2013;93(6):1151–5. http://dx.doi.org/10.1016/j.ajhg.2013.10.015.
[45] Gusev A, Bhatia G, Zaitlen NA, Vilhjálmsson BJ, Diogo B, Stahl EA, et al. Quantifying missing heritability at known GWAS loci. PLoS Genet 2013;9(12), e1003993. http://dx.doi.org/10.1371/journal.pgen.1003993.
[46] Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee B, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet 2015;47(10):1114–20. http://dx.doi.org/10.1038/ng.3390.
[47] Lee SH, Yang J, Goddard ME, Visscher B, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 2012;28(19):2540–2. http://dx.doi.org/10.1093/bioinformatics/bts474.
[48] Vattikuti S, Guo J, Chow CC. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet 2012;8(3), e1002637. http://dx.doi.org/10.1371/journal.pgen.1002637.
[49] Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet 2013;45(9):984–95. http://dx.doi.org/10.1038/ng.2711.
[50] Lee SH, DeCandia TR, Ripke S, Yang J, Schizophrenia Psychiatric Genome-Wide Association Study Consortium, International Schizophrenia Consortium, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet 2012 b;44(3):247–50. http://dx.doi.org/10.1038/ng.1108.
[51] Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet 2014;95(5):535–52. http://dx.doi.org/10.1016/j.ajhg.2014.10.004.
[52] Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015; 47(3):291–5. http://dx.doi.org/10.1038/ng.3211.
[53] Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 2015. http://dx.doi.org/10.1038/ng.3404.
[54] Bulik-Sullivan B, Finucane B, Anttila V, Gusev B, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet 2015. http://dx.doi.org/10.1038/ng.3406.
[55] Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55(4): 997–1004. http://dx.doi.org/10.1111/j.0006-341X.1999.00997.x.
[56] Vattikuti S, Lee JJ, Chang CC, Hsu SDH, Chow CC. Applying compressed sensing to genome-wide association studies. GigaScience 2014;3:10. http://dx.doi.org/10.1186/2047-217X-3-10.
[57] Donoho DL, Maleki A, Montanari A. Message-passing algorithms for compressed sensing. Proc Natl Acad Sci U S A 2009;106(45):18914–9. http://dx.doi.org/10.1073/pnas.0909892106.
[58] Candès EJ, Romberg B, Tao T. Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math 2006;59(8):1207–23.
[59] Candès EJ, Plan Y. A probabilistic and RIPless theory of compressed sensing. IEEE Trans Inf Theory 2011;57(11):7235–54. http://dx.doi.org/10.1109/TIT.2011.2161794.
[60] Friedman J, Hastie T, Höfling H, Tibshirani B. Pathwise coordinate optimization. Ann Appl Stat 2007;1(2):302–32. http://dx.doi.org/10.1214/07-AOAS131.
[61] Chang CC, Chow CC, Tellier LCAM, Vattikuti B, Purcell SM, Lee B. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience 2015;4:7. http://dx.doi.org/10.1186/s13742-015-0047-8.
[62] Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet 2015;11(4), e1004969. http://dx.doi.org/10.1371/journal.pgen.1004969.
[63] Palla L, Dudbridge F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. Am J Hum Genet 2015;97(2):250–9. http://dx.doi.org/10.1016/j.ajhg.2015.06.005.
[64] Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet 2014;94(4):559–73. http://dx.doi.org/10.1016/j.ajhg.2014.03.004.
[65] Kichaev G, Pasaniuc B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. Am J Hum Genet 2015;97(2):260–71. http://dx.doi.org/10.1016/j.ajhg.2015.06.007.