RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Fine-Mapping Additive and Dominant SNP Effects Using Group-LASSO and Fractional Resample Model Averaging

Jeremy Sabourin,[1,2] Andrew B. Nobel,[2,3,4] and William Valdar[1,2]*

[1]*Department of Genetics, University of North Carolina at Chapel Hill, North Carolina, United States of America;* [2]*Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, North Carolina, United States of America;* [3]*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, North Carolina, United States of America;* [4]*Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, United States of America*

Received 20 May 2014; Revised 25 September 2014; accepted revised manuscript 30 September 2014.
Published online 21 November 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21869

ABSTRACT: **Genomewide association studies (GWAS) sometimes identify loci at which both the number and identities of the underlying causal variants are ambiguous. In such cases, statistical methods that model effects of multiple single-nucleotide polymorphisms (SNPs) simultaneously can help disentangle the observed patterns of association and provide information about how those SNPs could be prioritized for follow-up studies. Current multi-SNP methods, however, tend to assume that SNP effects are well captured by additive genetics; yet when genetic dominance is present, this assumption translates to reduced power and faulty prioritizations. We describe a statistical procedure for prioritizing SNPs at GWAS loci that efficiently models both additive and dominance effects. Our method, LLARRMA-dawg, combines a group LASSO procedure for sparse modeling of multiple SNP effects with a resampling procedure based on fractional observation weights. It estimates for each SNP the robustness of association with the phenotype both to sampling variation and to competing explanations from other SNPs. In producing an SNP prioritization that best identifies underlying true signals, we show the following: our method easily outperforms a single-marker analysis; when additive-only signals are present, our joint model for additive and dominance is equivalent to or only slightly less powerful than modeling additive-only effects; and when dominance signals are present, even in combination with substantial additive effects, our joint model is unequivocally more powerful than a model assuming additivity. We also describe how performance can be improved through calibrated randomized penalization, and discuss how dominance in ungenotyped SNPs can be incorporated through either heterozygote dosage or multiple imputation.**
Genet Epidemiol 39:77–88, 2015. Published 2015 Wiley Periodicals, Inc.*

KEY WORDS: **SNP prioritization; GWAS; dominance; variable selection; model averaging**

Introduction

Genomewide association studies (GWAS) have been highly successful at identifying short chromosomal regions (loci) harboring causal variants affecting complex disease. At these identified loci, significant associations of single-nucleotide polymorphisms (SNPs) with the trait, as judged by single-marker testing, are often numerous; yet in many cases, the pattern of association, considered jointly with inspection of the linkage disequilibrium (LD) structure, clearly implicates a few SNPs (genotyped or imputed) that best represent the underlying causal variants [Edwards et al., 2013]. In other cases, however, the pattern of association and LD structure is so complex that both the number and identities of such best representative SNPs are ambiguous [e.g., Strange et al., 2010]; such ambiguity is problematic because it supplies a poorly informed starting point for subsequent experimental or annotation-based followup.

At these more complex loci, there is a compelling case for reanalysis using procedures that model multiple SNP effects simultaneously. In theory, such multi-SNP methods should lead not only to more robust estimates of each SNP's effect but also should better distinguish which SNPs represent independent signals worthy of subsequent prioritization [Stephens and Balding, 2009; Wang et al., 2012]. In practice, however, the multi-SNP analysis typically performed is of an informal nature, restricted to, for example, conditional regressions on top-scoring associations, and producing statistics that, though superficially decisive (e.g., declaring the presence or absence of multiple signals), are also unreliable, in that they would be expected to vary dramatically in different samples from the same population (see well-established arguments in, e.g., Buckland et al., 1997). Although more sophisticated multi-SNP methods based on general statistical procedures for variable selection can provide more robust or moderated inference, these methods are much less often applied and their results less often reported. This may be in part because such methods, in balancing a more complicated set of statistical priorities, make trade-offs between computational

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: William Valdar, Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: william.valdar@unc.edu

convenience and biological comprehensiveness that can seem severe, arbitrary, and/or omitting of important features.

One biological feature routinely omitted in methods currently proposed for multi-SNP analysis is genetic dominance. Instead, it is common to assume additive-only genetics—that is, where the effect of each SNP's minor allele is strictly additive in relation to its count [Ballard et al., 2010; Kooperberg et al., 2010; Segura et al., 2012]. This assumption of additivity is likely motivated by convenience: Although modeling both additive and dominance effects together in the single-marker setting is straightforward [Yeager et al., 2007; Li et al., 2009; Zheng et al., 2011], in the multi-SNP setting it can be hard to achieve gracefully. In particular, each SNP is ideally represented by a pair of variables, but these variables are often highly correlated with each other. For reasons of speed or simplicity, the burden of these additional complications typically leads to dominance being ignored [see, e.g., Guan and Stephens, 2011].

The presence of genetic dominance, or equivalently genetic recessivity, is nonetheless a standard assumption made when modeling complex traits in plants and model organisms [Lynch and Walsh, 1998; Neale et al., 2008]. Dominance is also commonly recognized in human genetics: it is a well-studied feature of monogenic disorders such as cystic fibrosis and phenylketonuria; it has been shown to contribute significantly to broad-sense heritability in cohort studies [Ober et al., 2001]; and it has been shown to have a substantial role in the form of recessivity leading to inbreeding depression in height [McQuillan et al., 2012]. Moreover, dominance has been shown to capture with greater accuracy (and higher statistical significance) the effects at some disease susceptibility loci in human GWASs of blood pressure [Org et al., 2009], cancer [Antoniou et al., 2009], and kidney disease [Okamoto et al., 2011].

Modeling that ignores dominance will in many cases provide inference that is nonetheless adequate. For instance, when the local genetic architecture is in fact mostly additive; or when, in the sample under study, dominant-acting loci are anyway observed on only two genotypes (and thus well-modeled by an additive effect). Historically, assuming additivity even in single-marker tests was a reasonable trade-off when primarily using tagging SNPs: In that setting, dominant-acting causal SNPs can present an additive-like signal if and when detected through a tagging SNP in low to moderate LD [Chapman and Clayton, 2007].

In modern genetic association, however, with genotypes available at high density, and when strong genetic dominance is present, modeling effects at single SNPs under additive-only assumptions can fail badly [Kim et al., 2010]. By implication, when applied in this context, multi-SNP models that fail to capture dominance can give more weight to weak additive associations modeled well than to strong dominance signals modeled poorly; and when deciding among candidates for further investigation, this risks systematic prioritization of the wrong SNPs.

In this article we propose a multi-SNP method that uses joint modeling of additive and dominance effects to characterize more accurately the genetic architecture of loci identified in standard GWAS, producing a reprioritization of SNPs that is enriched for true signals relative to single-marker and additive-only multi-SNP analyses. Our method builds on our previous work in this area: We take our existing multi-SNP method LLARRMA [Valdar et al., 2012] (reviewed in Methods), which reprioritizes SNPs under the assumption of additive-only genetics, and we introduce several coordinated extensions that allow it to model dominance. The original LLARRMA combined LASSO-penalized regression with resampling in order to estimate how frequently each SNP at a locus would be included in a sparse multi-SNP model in a random sample from the population. This led to a prioritization of SNPs based on the stability of each SNP's association with the phenotype, with stability judged with regard to both sampling variation and competition from other SNPs that may be in high LD. In this article, we incorporate dominance through two (necessarily linked) elaborations: (1) each SNP, previously represented by a single additive-effect predictor, is now modeled as a pair of variables subject to joint shrinkage and selection via a group LASSO penalty; and (2) the resampling procedure, previously subsampling, is generalized to ensure that the dimension of the predictor matrix, and thereby the dimensions of the fitted LASSO model, between resamples remains constant. This latter elaboration, which we term "fractional resampling" and which is related to the Bayesian bootstrap [Rubin, 1981], helps resolve a general problem with the application of resampling-based methods to datasets in which some levels of a categorical predictor have low frequency—that is, where some genotypes necessary for defining a modeled effect are rare to the point of being entirely absent in a large proportion of ordinary resamples. Additionally, we investigate the use of randomized LASSO penalization [after Meinshausen and Bühlmann, 2010] as a means to stabilize behavior of the LASSO when applied to subsets of SNPs in extremely high LD, incorporating this feature into our procedure as an option that, once calibrated to the data under study, can provide inference that is even more robust.

Methods

We start by describing a standard linear model for estimating the additive and dominance (recessive) effects of m SNPs at a broad genomic locus (hereafter, "locus") identified by GWAS of a quantitative outcome in n individuals; we then describe a statistical procedure to identify a subset of m_q SNPs that might best represent the underlying causal variants using penalized regression combined with a resampling procedure. For convenience we define a "true signal" as the SNP that most strongly tags an underlying causal variant, a "background" SNP to be an SNP that is not a true signal, and an optimal analysis as one that distinguishes true signals from background SNPs within a locus. Throughout, we assume the following setup: that the locus has been previously identified by an initial genomewide scan (e.g., using single-marker regression); that the m SNPs may be in high LD; and that $m_q < m < n$. Notationally, we represent the true

(but unknown) status of the SNPs using the binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$, where $\gamma_j = 1$ if SNP j is a true signal and $\gamma_j = 0$ otherwise.

Statistical Model

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the n-vector of quantitative outcomes for the n individuals, and define the "additive" matrix $\mathbf{A}$ as an $n \times m$ matrix containing the minor allele count for each of the m genotyped SNPs—that is, for a given SNP j with alleles q (major allele) and Q (minor allele), the ith element a_{ij} is 0, 1, and 2 for genotypes qq, qQ, and QQ, respectively. The "dominance" matrix $\mathbf{D}$ is defined to be another $n \times m$ matrix indicating for each SNP j whether the genotype of individual i is a heterozygote; in particular, $d_{ij} = I_{\{a_{ij}=1\}}$ (as in, e.g., Servin and Stephens, 2007). The phenotype y_i of individual i is modeled by a linear regression of the $2m$ predictors as follows:

$$y_i = \mu + \mathbf{a}_{i,*}^{\mathrm{T}}\boldsymbol{\beta}_a + \mathbf{d}_{i,*}^{\mathrm{T}}\boldsymbol{\beta}_d + \varepsilon_i, \tag{1}$$

where μ is the intercept; $\mathbf{a}_{i,*}$ and $\mathbf{d}_{i,*}$ are the ith rows of $\mathbf{A}$ and $\mathbf{D}$, respectively; $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_d$ are the corresponding m-vectors of additive and dominance effects, respectively; and $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.

General Penalized Likelihood Framework.

Our primary goal is not to distinguish which SNPs are additive or dominant but rather to identify the m_q true signal SNPs, regardless of their genetic mode of action. For SNPs selected to be included in the model, it is then a secondary goal to estimate additive and dominance components of their effects. In order to achieve these goals, we adopt penalized likelihood approach that adapts ideas from the Bayesian bootstrap [Rubin, 1981], the group LASSO [Yuan and Lin, 2006], and the randomized LASSO [Meinshausen and Bühlmann, 2010]. Details of the penalized likelihood framework are given below; further discussion and motivation follow.

Let $\mathcal{D} = \{\mathbf{y}, \mathbf{A}, \mathbf{D}\}$ represent the observed data, and let $\boldsymbol{\beta}^{\mathrm{T}} = [\boldsymbol{\beta}_a^{\mathrm{T}}, \boldsymbol{\beta}_d^{\mathrm{T}}]$ be the vector of additive and dominance SNP effects. Let the vector $\boldsymbol{\theta} = (\mu, \sigma^2)$ collect all nuisance parameters and, optionally, any covariate effects. For $1 \le i \le n$, let $L(\boldsymbol{\beta}, \boldsymbol{\theta}; y_i, \mathbf{a}_{i,*}, \mathbf{d}_{i,*})$ be the likelihood of observation y_i given parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and genotype predictors $\mathbf{a}_{i,*}$ and $\mathbf{d}_{i,*}$. Given a sample weight vector $\mathbf{w} = (w_1, \ldots, w_n)$ with $w_i \in [0, \infty)$, define the weighted log-likelihood as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{w}, \mathcal{D}) = \sum_{i=1}^{n} w_i \log L(\boldsymbol{\beta}, \boldsymbol{\theta}; y_i, \mathbf{a}_{i,*}, \mathbf{d}_{i,*}). \tag{2}$$

When all weights w_i are equal to 1, the weighted log-likelihood reduces to the ordinary log-likelihood; for binary weights $w_i \in \{0, 1\}$, the weighted log-likelihood is the log-likelihood of a subsample consisting of those observations i for which $w_i = 1$; for normally distributed traits, Equation (2) simplifies to a weighted residual sum of squares.

In real data, for a given SNP j, the additive and dominance vectors $\mathbf{a}_{*,j}$ and $\mathbf{d}_{*,j}$ (i.e., the jth columns of $\mathbf{A}$ and $\mathbf{D}$, respectively) will often be highly correlated. To avoid this correlation producing numerically unstable results, it is advantageous to moderate the estimation of each nonzero pair $\{\beta_{a,j}, \beta_{d,j}\}$ through some form of joint shrinkage. We achieve this by structured penalization based on the group LASSO, which forces prespecified groups of parameters to enter or leave the model together, and performs shrinkage of effects within each group. In more detail, we employ a weighted penalty of the form

$$\mathrm{pen}(\boldsymbol{\beta}; \mathbf{r}) = \sum_{j=1}^{m} r_j^{-1}\sqrt{\beta_{a,j}^2 + \beta_{d,j}^2}, \tag{3}$$

where $r_1, \ldots, r_m \in (0, \infty)$ are SNP-specific weights.

Describing both observation-specific and SNP-specific weights collectively as a "perturbation," $\mathcal{R} = \{\mathbf{w}, \mathbf{r}\}$, and letting $\lambda > 0$ be a parameter balancing the tradeoff between the log-likelihood and the penalty, our estimate of additive and dominance effects is given by

$$\hat{\boldsymbol{\beta}}(\lambda, \mathcal{R}; \mathcal{D}) = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ -\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{w}, \mathcal{D}) + \lambda \cdot \mathrm{pen}(\boldsymbol{\beta}; \mathbf{r}) \right\}. \tag{4}$$

When all values in perturbation $\mathcal{R}$ are equal to 1, the estimate $\hat{\boldsymbol{\beta}}(\lambda, \mathcal{R}; \mathcal{D})$ coincides with that of the group LASSO [Yuan and Lin, 2006] with a natural grouping of additive and dominance effects for each SNP.

Permutation-Based Selection of the Penalty Parameter.

The penalty parameter λ in Equation (4) controls the degree of sparsity (and shrinkage) in the model, with larger values causing a greater number of SNPs to be excluded. To choose a suitable value of λ for inference, we use the discovery-based "permutation selection" criterion proposed in Valdar et al. [2012], which was based on a proposal of Ayers and Cordell [2010]. Let $\boldsymbol{\pi}_1(\mathbf{y}), \ldots, \boldsymbol{\pi}_S(\mathbf{y})$ be S permutations of the response $\mathbf{y}$, and define $\boldsymbol{\pi}_s(\mathcal{D}) = \{\boldsymbol{\pi}_s(\mathbf{y}), \mathbf{A}, \mathbf{D}\}$ be a version of $\mathcal{D}$ in which the response $\mathbf{y}$ is replaced by its sth permutation. For a given perturbation $\mathcal{R}$, let λ_s be the smallest value of penalty λ for which each entry of $\hat{\boldsymbol{\beta}}(\lambda, \mathcal{R}; \boldsymbol{\pi}_s(\mathcal{D}))$ is zero. The permutation selection value of λ is defined as the median of the λ_s values,

$$\hat{\lambda}(\mathcal{R}, \mathcal{D}) = \mathrm{Median}(\lambda_1, \lambda_2, \ldots, \lambda_S.). \tag{5}$$

For stable estimation of this median, we find that in real data settings $S = 20$ (as in Valdar et al., 2012) is adequate, and we use that value here.

Definition of an Included SNP.

For a fixed perturbation $\mathcal{R}$, the jth SNP is defined as being "included" if either its additive or dominant effect is estimated to be nonzero. Formally, if $\hat{\beta}_{a,j}$ and $\hat{\beta}_{d,j}$ are the additive and dominance effects for SNP j estimated in $\hat{\boldsymbol{\beta}}(\lambda, \mathcal{R}; \mathcal{D})$ where

$\lambda = \hat{\lambda}(\mathcal{R}, \mathcal{D})$, then we define the inclusion indicator of SNP j to be

$$\hat{\gamma}_j(\mathcal{R}, \mathcal{D}) = \begin{cases} 0 & \text{if } \hat{\beta}_{a,j} = \hat{\beta}_{d,j} = 0 \\ 1 & \text{otherwise.} \end{cases} \quad (6)$$

Calculating this for SNPs $j = 1, \ldots, m$ leads to the binary m-vector of SNP inclusions $\hat{\boldsymbol{\gamma}}(\mathcal{R}, \mathcal{D}) = (\hat{\gamma}_1(\mathcal{R}, \mathcal{D}), \ldots, \hat{\gamma}_m(\mathcal{R}, \mathcal{D}))$, which estimates the true vector of inclusions $\boldsymbol{\gamma}$.

Fractional Resample Model Averaging

Model Averaging Rationale.

Frequentist methods for variable selection commonly used in this context, including stepwise selection, LASSO, and informal conditional regression, typically yeild a single, categorical decision about which SNPs should be included in the model, that is, a hard estimate of $\boldsymbol{\gamma}$; but a statement of that estimate's frequentist properties, namely how much it would be expected to vary in alternate, random, finite samples from the population (i.e., akin to a standard error), is analytically unavailable and so goes unreported. This is unfortunate because the set of SNPs selected, along with any inferences conditioning on that set (e.g., the jointly considered SNP effects), is usually highly sensitive to sampling (see well-established arguments in, e.g., Buckland et al., 1997), and reporting it as if otherwise is at best optimistic, and at worst misleading. Faced with this, one option is to switch to using Bayesian methods, which explicitly model posterior uncertainty about higher order inference, following appropriate specification of priors (e.g., Guan and Stephens, 2011); another, which we consider here, is to stay within the frequentist framework but estimate relevant frequentist properties of the selection procedure by means of resampling.

Rather than using LASSO-based model selection to estimate $\boldsymbol{\gamma}$ directly, we instead use multiple estimates of $\boldsymbol{\gamma}$ obtained from alternative realizations of the data to help identify those SNPs whose association with the phenotype is stable, both to model selection and sampling variation. In our previous work, these alternative datasets were generated by bootstrapping or subsampling the observations, and optionally using a multiple imputation step to fill in any missing data [Valdar et al., 2006, 2009, 2012]. That general approach we termed resample model average (RMA) to reflect the fact that it generalizes bagging [Breiman, 1996] and bootstrap model averaging [Buckland et al., 1997] (see also Utz et al., 2000). In RMA, statistics of interest are collected on each resample (and imputation) and then averaged (or otherwise summarized) to give a final, aggregate estimate. In the context of fine-mapping SNPs at a locus: for each SNP j we estimate the long-run average frequency $\mathrm{E}(\gamma_j)$ that its association with the phenotype would survive model selection, describing this as its resample model inclusion probability (RMIP; Valdar et al., 2009). In Valdar et al. [2012] we developed an RMA method based on the LASSO (termed LLARRMA, for LASSO local automatic regularization RMA) to reprioritize SNP associations at a locus (or "hit region," in that article's terminology) under an additive genetic model; here we extend that work to incorporate modeling of dominance.

Extending the additive model to incorporate dominance, however, poses challenges for RMA: Aggregating the results of model selection and estimation over repeated bootstraps or subsamples of data $\mathcal{D}$ is problematic when $\mathcal{D}$ contains predictors that are constant among a large fraction of individuals. For example, SNPs with a small number of individuals representing their homozygous minor allele genotype (i.e., QQ) will have indistinguishable additive and dominance components in some resamples but not others; rarer SNPs with few examples of variant individuals may become entirely monomorphic. In either case, SNPs with low frequencies for some genotypes will not be representable in the same way between resamples, leading to a change in dimensionality of $\boldsymbol{\beta}$ or $\hat{\boldsymbol{\gamma}}$ that complicates the definition and interpretation of statistics such as RMIPs that are based on aggregation.

Resample Model Averaging Using Fractional Observation Weights.

To overcome the problem of rare predictors, we introduce a generalization of RMA, fractional resample model averaging (FReMA), based on fractional data weights. Let $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in (0, \infty]^n$ be vectors of nonnegative sample weights drawn independently from a fixed weighting distribution $W(\cdot)$. For each $k = 1, \ldots, K$, define the following quantities: the kth perturbation $\mathcal{R}^{(k)} = \{\mathbf{w}^{(k)}, \mathbf{r}^{(k)}\}$, where $\mathbf{r}^{(k)} = \mathbf{1}_m$ unless otherwise stated; the penalty $\hat{\lambda}^{(k)} = \hat{\lambda}(\mathcal{R}^{(k)}, \mathcal{D})$; the vector of estimated effects $\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}(\hat{\lambda}^{(k)}, \mathcal{R}^{(k)}; \mathcal{D})$; and the vector of SNP inclusions $\hat{\boldsymbol{\gamma}}^{(k)} = \hat{\boldsymbol{\gamma}}(\mathcal{R}^{(k)}, \mathcal{D})$. Performing K independent fractional resamples gives the $m \times K$ matrix of inclusions $\boldsymbol{\Gamma} = [\hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\gamma}}^{(2)}, \cdots, \hat{\boldsymbol{\gamma}}^{(K)}]$. From this matrix, we estimate the fractional RMIP (FReMIP) of SNP j as

$$\widehat{\mathrm{FReMIP}}_j = \frac{1}{K}\sum_{k=1}^{K} \hat{\gamma}_j(\mathcal{R}^{(k)}, \mathcal{D}) = \frac{1}{K}\sum_{k=1}^{K} \hat{\gamma}_j^{(k)} = \frac{1}{K}\sum_{k=1}^{K} \Gamma_{jk}, \quad (7)$$

which approximates the expectation of γ_j over repeated samples from the same weighting distribution. Similarly, for any function $g(\boldsymbol{\beta})$, such as a predicted phenotype or an SNP effect, we can obtain an FReMA estimate as $\hat{g} = K^{-1}\sum_k g(\hat{\boldsymbol{\beta}}^{(k)})$.

Weighting distributions for which $w_i^{(k)} = 0$ with probability zero fix the dimensionality of the data, and thereby the dimensionality of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}}$, across resamples because they ensure that every observation participates (to a lesser or greater extent) in the likelihood. For normally distributed traits, sample weighting is equivalent to some phenotype values being treated as if they were measured with less precision, because the weighting in Equation (2) becomes equivalent to respecifying the residuals in Equation (1) as $\varepsilon_i \sim \mathrm{N}(0, \sigma^2/w_i)$.

Proposed Weighting Density: Independent Uniforms.

We propose to sample the weight of each data point independently from the range 0 to 1, that is, to draw $w_i^{(k)} \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. This can be seen as a smoothed version of subsampling. Under this setting, the effective sample size of each fractional resample, which we define to be the total weight $n^{(k)} = \sum_{i=1}^{n} w_i^{(k)}$, has expectation $\text{E}(n^{(k)}) = \frac{1}{2}n$ and, for reasonable values of n, has approximate distribution $\text{N}(\frac{1}{2}n, (12n)^{-1})$. Moreover, the proportion of total data weight $n^{(k)}$ attributed to any one observation will generally be of order $1/n$ with maximum proportion of order $2/n$.

Alternative Weighting Densities: Subsampling and Bootstrapping.

The weighting density W is defined loosely, and encompasses subsampling, nonparametric bootstrapping and the Bayesian bootstrap, among others. Specifically, when $W(\cdot)$ is generated by permuting ϕn ones and $(\phi - 1)n$ zeros for some proportion ϕ, then FReMA is equivalent to subsample model averaging (e.g., Valdar et al., 2006, 2012); when $W(\cdot) = \text{Multinomial}(n^{-1}\mathbf{1}_n, n)$, such that $w_i^{(k)} \in \{0, 1, \ldots, n\}$ subject to $n^{(k)} = n$, FReMA is equivalent to bootstrap model averaging (e.g., Buckland et al., 1997); and when $W(\cdot) = \text{Dirichlet}(\mathbf{1}_n)$, such that $w_i^{(k)} \in [0, n]$ subject to $n^{(k)} = n$, it is equivalent to model averaging based on the Bayesian bootstrap [Rubin, 1981; Clyde and Lee, 2001]. The maximum proportion of weight attributed to any one individual varies among these alternatives: for subsampling it is constant at $1/n$, but for the bootstrap and Bayesian bootstrap it is considerably greater (of order $6/n$ or more).

Terminology for Resampling.

Despite the fact that FReMA generalizes RMA, for the purposes of this article we will henceforth reserve the terms "fractional resample," FReMIP, and FReMA exclusively for procedures and outcomes that involve weighting distributions capable of producing noninteger weights (e.g., independent uniform weighting and the Bayesian bootstrap). For continuity with previous work, we will retain the terms RMA and RMIP when referring to subsampling or bootstrapping (i.e., integer weight resamples), and will use "resample" to mean an integer or noninteger weight resample.

Selecting Among Highly Correlated SNPs: The Randomized Group-LASSO

When the LASSO is applied to a set of highly correlated SNPs that all strongly associate with the phenotype, it will tend to select a small subset arbitrarily. The effect of this arbitrariness on inference is mitigated by resampling only somewhat. For instance, in the extreme case of high correlation in the population leading to some pairs of SNPs having a perfect correlation in the sample, subsequent resampling will simply replicate this perfect correlation, with any arbitrary selections between the two SNPs being similarly replicated.

A modification of the LASSO that helps address this issue was proposed by Meinshausen and Bühlmann [2010]. In their "randomized LASSO," the penalty applied to each predictor in each resample is chosen at random from $\{\lambda, \lambda/c\}$, for some prespecified $c \in [0, 1]$ (see also Bühlmann and van de Geer, 2011). Here we employ a "randomized group LASSO" in which SNP-specific weights $\mathbf{r}^{(k)} = (r_1^{(k)}, \ldots, r_m^{(k)})$ are drawn independently in each resample as $r_j = \{c, 1\}$ with equal probability. Incorporating this additional randomization step into our procedure has minor consequences for computation time (see Supplementary Material).

Table 1. Nomenclature for modeling and resampling procedures used in the article

Character	Description
r	Randomized LASSO penalization
s	Resampling by subsampling
w	Resampling by fractional subsampling
a	Additive effects
d	Dominance/heterosis effects
g	"a" and "d" modeled as a grouped effect

Empirical Tuning of the Randomization Penalty

We find that the optimal choice of randomization parameter c depends on the correlation structure of the data, and so we recommend dataset-specific calibration. Moreover, because of the additional variability induced in the estimation, we recommend a minimum of $K = 250$ resamples, with more preferred. Based on preliminary simulations using the locus dataset described below, we find that optimal performance is achieved with $c \in [0.6, 0.8]$ (see Supplementary Material); in the applications of the randomized LASSO described subsequently we therefore set $c = 0.7$.

LLARRMA-rdawg Nomenclature

To help assess how much each proposed extension to LLARRMA contributes to performance, we use the naming convention in Table 1, suffixing "LLARRMA" with a string of letters indicating which genetic model, weighting density, and effect penalization is in place. Thus, "LLARRMA-dawg" applies fractional resampling to additive and dominance (recessive) effects grouped through a penalty applied to each SNP, and "LLARRMA-rdawg" additionally incorporates penalty randomization to help break ties between tightly linked SNPs. By contrast, "LLARRMA-sa" corresponds to the original, additive-only subsampling procedure of Valdar et al. (2012).

Competitor Method: Single-Marker Regression

Our model averaging approach calculates a score (an RMIP, or its generalization, an FReMIP) that prioritizes each SNP

in the identified locus. We compare the ability of those scores to discriminate true signals from background SNPs, not only between different variations of LLARRMA, but also with the standard GWAS alternative: single-marker regression, as used in, for example, PLINK [Purcell et al., 2007]. In single-marker regression, we fit a linear model to each SNP with either additive-only or additive and dominance effects as in Equation (1), and report the score of the SNP as $-\log_{10} P$ (or logP), where P is the P-value from a likelihood ratio test against an intercept-only model. For comparison of LLARRMA-sa with multi-SNP procedures based on stability selection and Bayesian model averaging, we refer the reader to Valdar et al. [2012].

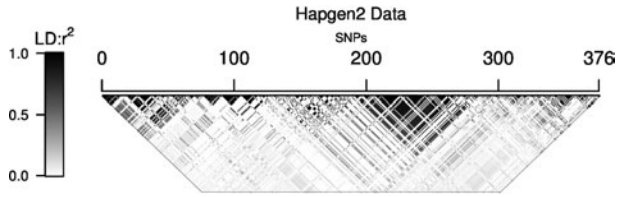

Figure 1. LD structure used by HAPGEN2 to create simulated genotype data. Shading indicates pairwise LD between SNPs, ranging from white ($r^2 = 0$) to black ($r^2 = 1$).

Simulation Framework

Simulation was used to assess how well each method could discriminate SNPs with additive and/or dominance effects from those with no effect in GWAS loci of complex LD. Performance was examined under seven different settings (locus effect architectures), each corresponding to a scenario in which a different mix of additive and dominance effects were present. Simulations were structured as follows: first, 500 loci, each comprising 376 SNPs on 2,500 individuals, were simulated based on real GWAS data; for each simulated locus, five SNPs were randomly selected to be true signals; true signals were then assigned effects according to the simulated effect architecture; the effects of the five true signals were combined with random noise to generate a quantitative phenotype owing on average 5.8% of its variance to the true signals; phenotype and genotype data were analyzed by LLARRMA, its extensions, and single-marker regression to produce scores (logP, RMIP, or FReMIP) for each SNP; and the abilities of these scores to discriminate the true signals from background SNPs were then assessed. Details of specific steps follow.

Genotypes.

Loci were simulated to have an LD structure mimicking that of a real reference dataset. The reference dataset in this case comprised genotypes for 2,016 individuals on 376 SNPs in the contiguous 39.063723–40.985321 Mb segment on chromosome 22 from the human GWAS of WTCCC [2007] (LD structure shown in Fig. 1, note: this is a trimmed version of the '58 data used in Valdar et al., 2012). Genotypes for new loci were generated as follows: haplotypes were inferred for each individual in the reference dataset using fastPHASE [Scheet and Stephens, 2006]; these haplotypes were passed to HAPGEN2 [Su et al., 2011], together with necessary map and legend files from the HapMap project [The International HapMap 3 Consortium, 2010] (specifically these files were from HapMap 3 [release 2] NCBI build 36 [dbSNP b126]); HAPGEN2 was then used to generate haplotypes and corresponding genotypes for 376 SNPs on 2,500 new individuals.

Placement of True Signals.

For each simulated locus, five SNPs $\mathcal{Q} = \{q_1, \ldots, q_5\} \subset \{1, \ldots, m\}$ were selected to be true signals. Selection was at random, subject to the minor allele frequency (MAF) of any chosen SNP being at least 0.1; this bounded the genotype frequency of the homozygous minor allele (required to model dominance) to 0.01, ensuring that estimation of dominance was meaningful.

SNP Effects and Locus Effect Architecture.

True signals were assigned effects based on the simulated locus effect architecture. Seven such architectures were considered: additive effects only (architecture A); dominance of minor alleles (B); dominance of major alleles (C); heterosis (D); a mixture of dominance types (E); and mixtures of additive and dominance that are mostly additive (F) or mostly dominant (G). Architectures were defined by the SNP effect types they included. Specifically, for a given locus/architecture combination, each true signal was designated one of five SNP effect types—additive, or one of four sub-categories of dominance (first six columns of Table 2)in a probabilistic manner, with the probabilities of each type being

Table 2. Additive and dominance effects for simulated true signals

SNP effect type	Effects parameters SNP q		Genotype value			Locus effect architecture						
	Additive	Dominance	qq	qQ	QQ	A	B	C	D	E	F	G
Additive	β_q	0	0	β_q	$2\beta_q$	1	-	-	-	-	0.6	0.3
Minor allele dominant	β_q	β_q	0	$2\beta_q$	$2\beta_q$	-	1	-	-	-	-	-
Major allele dominant	β_q	$-\beta_q$	0	0	$2\beta_q$	-	-	1	-	-	-	-
Heterosis	0	β_q	0	β_q	0	-	-	-	1	-	0.1	0.1
General dominance	β_q	$\delta\beta_q$	0	$(1+\delta)\beta_q$	$2\beta_q$	-	-	-	-	1	0.3	0.6

architecture-dependent (listed in the last seven columns of Table 2, where blank cells "-" denote probability zero). For example, under setting A only additive effects are assigned; under setting G additive effects are assigned with probability 0.3, with heterosis and general dominance assigned with probabilities 0.1 and 0.6, respectively. Once an effect type of some SNP q had been designated, its corresponding effects were assigned in a stochastic fashion: Additive effect $\beta_{a,q}$ and dominance effect $\beta_{d,q}$ were calculated according to the rules in columns 2 and 3 in Table 2, where, independently for each simulation trial and for each locus, we drew sign variable $B \sim \text{Bernoulli}(\frac{1}{2})$, dominance variable δ randomly from {–1.25, –1, –0.75, –0.5, 0.5, 0.75, 1, 1.25}, and effect variable $\beta_q \sim \text{N}(1.35(-1)^B, 0.02^2)$.

Phenotypes and Signal-to-Noise Ratio.

Phenotypes were simulated based on the linear model in Equation (1). For a prescribed set of SNP effects $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_d$, we calculate the expected phenotype $\text{E}(y_i)$ for each individual i and then add residual error $\varepsilon_i \sim \text{N}(0, \sigma^2)$ to obtain simulated phenotype y_i. The residual error variance σ^2 was chosen to produce an expected signal to noise ratio of $\sigma^{-2}\boldsymbol{\beta}^{\text{T}}\text{Var}([\mathbf{A}, \mathbf{D}])\boldsymbol{\beta} = 1/16$. This level of noise on average produced a locus explaining about 5.8% of the phenotypic variance, which is comparable to that observed in Warren et al. [2012] and Dastani et al. [2012].

Evaluating Performance: Initial ROC Curve and Initial AUC.

Methods were compared based otheir ability to discriminate true signals from background SNPs, with performance evaluated formally using receiver operator characteristic (ROC) curves. The way ROC curves are defined and used can vary between studies [Krzanowski and Hand, 2009]. Here, we follow the procedures and calculations described previously in [Valdar et al., 2012], which focus on the "initial ROC curve" and maximizing the corresponding initial area under curve (AUC): the "initial ROC curve" is defined as the segment of the ROC curve in which between 0% and 5% of the included SNPs are false positives; the "initial AUC" is the area underneath, proportionally rescaled to the range [0,1] such that an initial AUC of 1 is ideal whereas an initial AUC of 0.025 corresponds to selection based on random guessing. Examination of the initial ROC curve is, in our view, more relevant to GWAS practice than examination of the "full" ROC curve for the following reason: the initial AUC credits enrichment of true signals among only the top-scoring SNPs; unlike the full AUC, it gives no credit to, for example, a method that detects all true signals exclusively within the middle ranks of its discoveries.

Computation.

All analyses were performed in R [R Development Core Team, 2012], with the *glmnet* package [Friedman et al., 2010] used for fitting LASSO models and the *grplasso* package [Meier, 2009] for group LASSO models. For the purposes of the simulation study, the number of resamples used by each LLARRMA-based method was limited to $K = 250$.

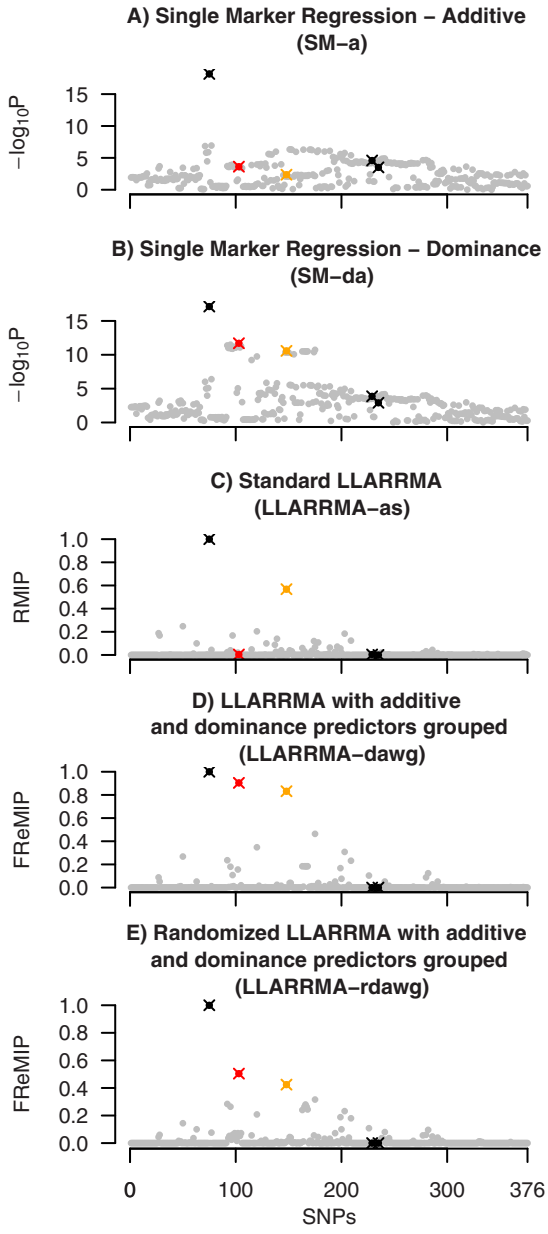

Figure 2. Results of five methods applied to an example dataset from simulation study G. Plots show SNP location in a simulated locus against SNP score, where SNP score is defined for single-marker methods (sub plots A, B) as $-\log_{10}$ of the *P*-value (logP), and for multi-SNP methods as the inclusion probability (RMIP or FReMIP). SNPs for which a true signal was simulated are shown as crosses: in black (additive effect), orange (major allele dominant), and red (minor allele dominant); all other (background) SNPs are shown as gray dots.

Results

Example output from a single simulation under setting G is shown in Figure 2. In this simulation, the locus includes both additive and dominance effect true signals. For

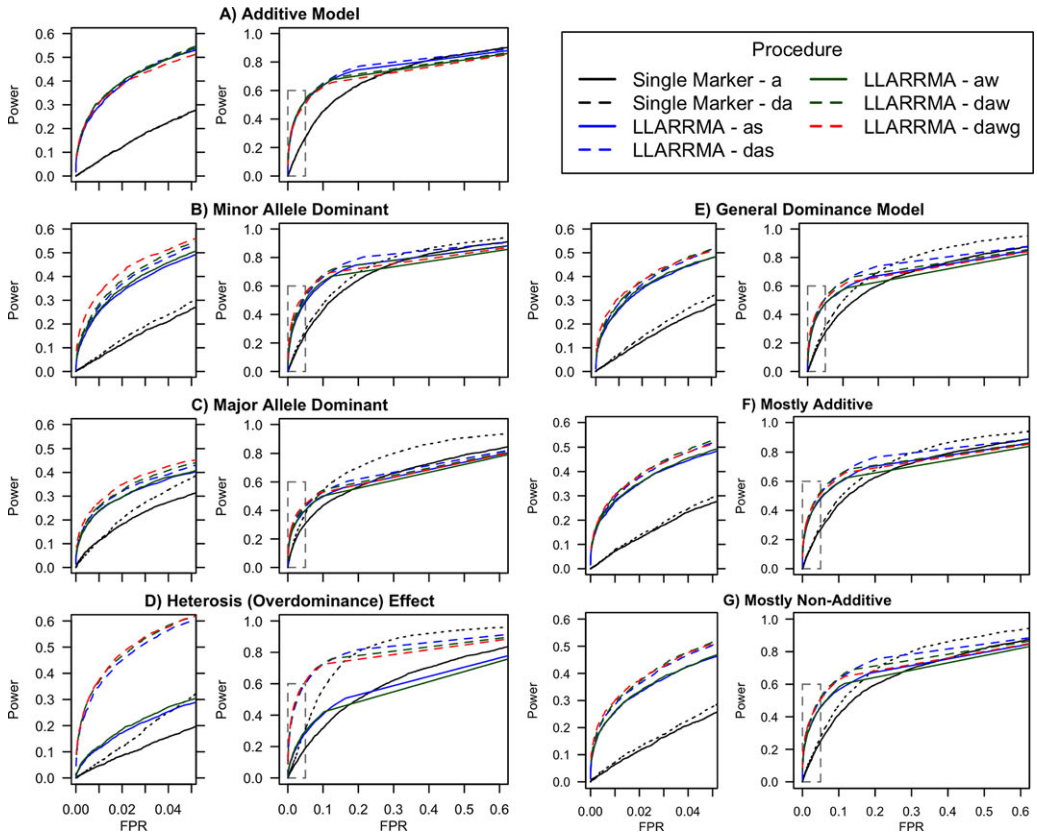

Figure 3. Initial and full ROC curves plotting discriminatory power of single-marker and multi-SNP (LLARRMA) methods under seven simulation settings. Each setting (A–G) corresponds to the presence of a different underlying genetic architecture for the locus, and plots both the full ROC curve (right) and its zoom-in (the initial ROC; left); better performance in each case is indicated by a greater area under the initial ROC curve. For clarity, only multi-SNP methods using constant penalization are shown.

five of the methods considered, Figure 2 plots SNP location against SNP score (logP, RMIP or FReMIP); true signals are shown as crosses colored black (additive effects), orange (major allele dominant), and red (minor allele dominant), with background SNPs in gray. Moving from left to right in the locus: the leftmost true signal, an SNP with an additive effect, is easily identified by all methods; the two dominance effect signals (red and orange) are given low priority by additive-only methods (SL-a and LLARRMA-as), but higher priority in methods that explicitly model dominance; the two rightmost true signals, additive effects that (at least in this example) are in repulsion, are essentially undetectable by any method. For this particular simulated locus, multi-SNP modeling (bottom three plots) is seen to act like a filter, assigning lower priorities to background SNPs that would otherwise detract attention from the true signals; in other examples, the nature of the reprioritization can be very different, and so general behavior of these methods is best considered through examining results from a large number of trials, as described below.

Loci With Additive-Only Effects

It was important for us to examine whether the incorporation of dominance effect parameters might lead to reduced power when dominance was absent. In Figure 3A, we therefore assess the performance of the LLARRMA variations against each other and against single-marker regression for additive only loci. Figure 3A contains two subplots: rightmost is the ROC curve plotting power (the proportion of true signals declared as influential) against false-positive rate (FPR; the proportion of background SNPs declared as influential) for each method over a range of possible cutoffs of the SNP score; leftmost is a zoom-in of the ROC curve for FPR ≤ 0.05, which we call the "initial ROC" curve. The initial ROC curve, and the area under it (the "initial AUC," rescaled to range [0,1]) are our primary performance measures (see Methods); values of the initial AUC are listed for each simulation in Table 3.

As shown in Figure 3A and Table 3, the LLARRMA-based methods are considerably more powerful than single-marker regression; but when compared with each other, at least for this additive-only locus, the LLARRMA-based methods are about evenly matched. Among methods using constant penalization, a slight advantage over other methods is seen for LLARRMA-daw (i.e., LLARRMA using fractional resampling and additional ungrouped predictors for dominance). Methods using randomized penalization performed consistently better, and among these the most successful were the

Table 3. Mean percentage of maximum initial AUC for all simulation studies (all standard errors < 0.26)

Locus effect architecture	Single-marker regression		LLARRMA Constant penalization					LLARRMA Randomized penalization				
(simulation study)	-a	-da	-as	-das	-aw	-daw	-dawg	-ras	-rdas	-raw	-rdaw	-rdawg
Additive (A)	14.6	14.6	39.1	39.4	39.8	40.1	38.5	41.2	41.3	**42.5**	**42.7**	41.4
Minor dom. (B)	13.8	15.6	34.4	37.4	35.6	38.7	42.0	36.3	39.1	38.0	40.9	**45.4**
Major dom. (C)	19.4	22.3	30.1	32.1	30.4	32.9	34.8	32.0	33.2	32.5	34.4	**36.9**
Heterosis (D)	10.1	15.4	18.3	45.3	19.5	47.0	47.4	19.1	46.5	20.8	49.1	**50.3**
General dom. (E)	14.9	17.1	35.1	37.4	35.6	38.2	38.8	37.4	39.7	38.3	40.8	**42.5**
Mostly add. (F)	14.9	15.9	35.7	38.3	36.2	39.1	38.7	37.7	40.0	38.5	**41.0**	**41.3**
Mostly nonadd. (G)	13.5	15.0	33.8	37.2	34.0	38.0	38.3	35.7	39.0	36.5	40.2	**41.3**

For each simulation study, best performing method (or methods when other methods are not statistically different from the best method) in bold; best among those with constant penalization are underlined.

ungrouped, FReMA-based LLARRMA-raw and LLARRMA-rdaw. Results under constant penalization have been separated from those under randomized penalization to reflect the fact that the latter require an additional, potentially arduous data-dependent calibration (see Methods). Under the additive-only setting described here, group penalization conferred little advantage nor disadvantage to discriminatory power.

Loci With Dominance-Only Effects

When the simulated locus contained only dominance signals, best performance in the initial AUC was seen for multi-SNP methods using the grouped penalty (Figure 3 B–E and Table 3), specifically: LLARRMA-dawg among methods using a constant penalization, LLARRMA-rdawg among methods using randomized penalization, and LLARRMA-rdawg over both types of penalization (Table 3). The degree of outperformance varied with the type of dominance simulated, with the most extreme form—heterosis (simulation D)—producing the most striking contrasts: Under heterosis, additive-only models failed badly; adding a dominance term, especially in conjunction with fractional resampling, rescued performance (Table 3, Figure 3 D). More generally, simulations B–E revealed several clear trends: single-marker regression with additive-only effects is inferior to single-marker regression incorporating dominance; this in turn, perhaps surprisingly, is inferior to multi-SNP methods with additive-only effects; those in turn (less surprisingly) are inferior to multi-SNP methods incorporating dominance. Among multi-SNP methods, the incorporation of grouped penalization, randomized penalization, or (ideally) both, leads to a small but consistent performance increment.

Identifying multiple true signals in a locus of complex LD inevitably requires examining some number of false positives; superior analysis methods will be those for which that number is typically few. For each analysis method, we calculated the average number of SNPs that would need to be examined in order to find k true signals. This average is plotted in Figure 4 for $k = 1, \ldots, 4$, with results shown for single-marker regression and the constant-penalty LLARRMA methods. (Results for detecting the full $k = 5$, and those using randomized penalization, are omitted for legibility). When it comes to detecting the first ($k = 1$) of five true signals, we observe

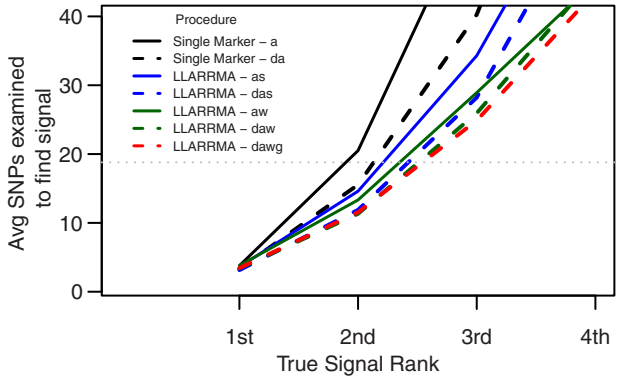

Figure 4. The average number of SNPs that must be examined to find the first, second, third, fourth, and fifth true signal in simulation E. Dotted gray line indicates 5% of the SNPs in the locus.

little difference between the methods; for detecting the second, third, and fourth ($k = 2, \ldots, 4$) true signals, single-marker regression quickly falls behind the multi-SNP methods, of which the FReMA-based, group-penalized LLARRMA-dawg is seen to be slightly more efficient for these simulations.

Loci With Additive and Dominance Effects

Under the potentially more common setting of both additive-only and dominance effect SNPs, results were more consistent with the simulations of dominance (B–E) than of additive-only loci (A): Incorporating dominance effects in the model significantly improved performance; the use of grouped and randomized penalties also each produced a small but consistent improvement. In particular, when true dominance effects exist, incorporating dominance in the model was most effective when modeled with additive effects as a group.

Discussion

We describe a statistical model selection procedure that better characterizes the genetic architecture at loci identified in GWAS by exploiting both additive and dominance components of each SNP's genetic effect. We show that when

it comes to discriminating additive-only true signals from background SNPs, there is no disadvantage associated with including dominance parameters, at least in our penalized regression framework; when, on the other hand, dominance effects are actually present—even in combination with significant additivity—we find that modeling both additive and dominance, either through a group penalty or as independently penalized predictors, produces a substantial improvement in discriminatory power.

The fact that additional power would be gained through more flexible or comprehensive modeling of effects at individual loci should not, in principle, be surprising: Dominance (or recessivity) is defined by its deviation from additivity; the more dominance signal present, the less effectively it will be captured by an additive model. In older GWAS focusing on indirect association through tagging SNPs, incorporating dominance effects adds little value because dominance, when identified by proxy through variants in low to moderate LD, resembles an additive signal [Chapman and Clayton, 2007]. In modern settings, however, when genotypes are available at higher density (or in the limit all variants are observed or well-imputed), where dominance effects are not attenuated in this way, and where complex LD can result in a locus showing multiple variants of similar significance, comprehensive modeling is more crucial. In our multi-SNP setting, where SNPs are reprioritized on the basis of their ability to compete with each other for a position in a sparse multi-SNP model, additive-only modeling gives dominance SNPs a competitive disadvantage; as shown in our simulations, this translates to poorer discrimination of true signals when dominance is present to any degree.

Dominance is modeled in our regression setting using a standard parameterization based on orthogonal contrasts that splits each biallelic effect into separate additive and dominance components (also used by, e.g., Servin and Stephens, 2007). An alternative parameterization, also standard in genetic analysis, is based on ANOVA-style treatment contrasts whereby each genotype is modeled as a treatment category with its own effect, with one of these categories (usually homozygote qq) chosen as a baseline (e.g., Yang et al., 2010). The choice of parameterizing one way or the other does not affect how well a single- or multi-SNP model fits in terms of its likelihood; they are equivalent in this respect, and equivalently powerful. The choice does matter, however, when these coefficients representing the SNP effect are estimated subject to a penalty, such as the LASSO or group LASSO. In simulations reported in Supplementary Material, we show that LLARRMA-dawg using the orthogonal additive/dominance parameterization outperforms the genotype category parameterization in six of the seven genetic architectures tested here, with the degree of outperformance in five of these being dramatic. As discussed in the context of Bayesian priors by Servin and Stephens [2007], penalization imposes constraints that can strongly favor the fit of one parameterization over the other; under group LASSO penalization, we find that the orthogonal additive/dominance split used here confers substantially greater power and discrimination.

Despite its potential value, incorporating dominance into the fine-scale analysis of GWAS loci will likely require a slight shift in the use of results from SNP imputation. Imputation is a stalwart of GWAS, used to estimate genotypes of SNPs that either failed on the array or that were not genotyped explicitly, and thereby expanding considerably the set of SNPs available for identifying associations. The underlying statistical models used to perform the imputation, by their construction, distinguish all three genotypic classes of a SNP, and give probabilities for each; but in practice, information on these three genotypic classes is not fully exploited by subsequent analyses. In particular, the imputation of an SNP, which estimates (at least) three probabilities ($\mathbf{p} = \{p_{qq}, p_{qQ}, p_{QQ}\}$) that correspond to two nonredundant parameters, is most often reduced to a single value: an allele dosage ($p_{qQ} + 2p_{QQ}$). Although this allele dosage is intuitive and simple to incorporate into analyses, it represents a loss of information: The SNP effect is represented only in its additive component, whereas the dominance component (e.g., representable as heterozygote dosage p_{qQ}), regardless of whether it would have provided important signal, is discarded. For modeling dominance with imputed SNPs in LLARRMA (and its variations), we offer two options: (1) additive and dominance inputs can be substituted by their dosage and heterozygote dosage (as above) respectively; or, our preferred option when feasible, (2) in each resample, genotypes of imputed SNPs can be directly sampled, either as multinomial draws from their marginal probabilities $\mathbf{p}$ or, better, from the full phased haplotype-based posterior (as available in, e.g., Scheet and Stephens, 2006, and performed in Servin and Stephens, 2007 and Valdar et al., 2012).

In reporting results from our reprioritization methods, we have focused not on identifying a single, best, multi-SNP model, which under realistic conditions would be highly sample-dependent, but rather on estimating for each SNP a measure that reflects the robustness of its association, both to finite sampling and competition from alternative multi-SNP models. Our (fractional) resample model inclusion probabilities—RMIPs or FReMIPs—are superficially analogous to the posterior model inclusion probabilities generated by Bayesian variable selection procedures, but in fact have a wholly frequentist motivation: They reflect variability in an estimator with respect to finite sampling, where the estimator in this case is our group LASSO model and where sample-to-sample variability is emulated via resampling of the original data. The role played by resampling is therefore crucial to the practical utility of our approach.

In our previous resample model averaging procedure, LLARRMA, in which SNPs were modeled under additive-only assumptions [Valdar et al., 2012], resampling was performed using subsampling. When extending LLARRMA to model dominance, however, we encountered a shortcoming of resampling when applied to model selection on datasets in which some predictors have categories that are rarely observed: When SNPs have some genotype classes observed at low (but nonzero) frequencies, subsampling (or equivalently bootstrapping) can lead to a reduction in the number of

predictors associated with its effect. This means that rare variants can become monomorphic, and three-genotype SNPs, for which additive and dominance effects are distinguishable, can be converted into two genotype SNPs, for which additive and dominance cannot be separated. Although within a resample this change in effective dimension can be addressed by simply reducing the set of predictors from which included loci are chosen, when aggregating fitted models across resamples, it creates an additional level of incomparability that we find undesirable. Specifically, we aim to approximate the results of a hypothetical study that quantifies how much the sparse subset selected from a fixed (and nonsparse) candidate set of SNP predictors will vary with different samples of individuals drawn from the same population; but inducing variability in both the selection and the candidate set from which selection is made makes the resulting inclusion probabilities harder to interpret. To address this issue, we develop a smooth generalization of the resampling procedure itself—fractional resampling—that allows the candidate set to stay constant, permitting multi-SNP additive and dominance modeling even when the minor allele frequencies of those SNPs are low.

We find in simulation that using fractional resampling in place of subsampling leads to a consistent increase in discriminatory power; this is not only in settings where subsampling is likely to alter the candidate set (i.e., when modeling additive and dominance) but also in settings where subsampling should be unproblematic (i.e., when applying additive-only multi-SNP models). Our generalization of subsampling to fractional resampling is similar in spirit to the generalization of the nonparametric bootstrap to the Bayesian bootstrap [Rubin, 1981]: In the Bayesian bootstrap, the bootstrap's discrete multinomial weights are smoothed to a continuous dirichlet weighting; in fractional resampling, the effective 0 or 1 weights of subsampling are smoothed to fractions drawn from a uniform distribution on [0,1]. Moreover, both can be formulated as specific examples of the weighted likelihood bootstrap [Newton and Raftery, 1994]. Given the choice, however, we prefer our subsampling-based strategy over the Bayesian bootstrap because the latter allows a potentially much larger number of duplicates of some observations—a feature that we find hard to justify in the human genetics context (and a feature also criticized at length in Rubin, 1981). On the other hand, if subsampling can be viewed as approximately like drawing from Beta(a, a) where $a = 1/\infty$, then fractional resampling as we implement it here (i.e., $a = 1$) lies somewhere between subsampling and the application of a constant half-weight (i.e., $a = \infty$); thus despite the strong empirical performance of using Beta(1, 1), we acknowledge that other ways to generate fractional weights are also worth investigating.

Our results suggest that, when it comes to discriminating multiple true signals among highly correlated background SNPs, performance of the LASSO-based RMA procedures can be consistently improved by randomization of the penalty parameter. This randomization, whereby in each resample approximately half of the predictors are given a weakened penalty, acts to break ties and redistribute advantages between SNPs in extremely high LD; in doing so, it helps compensate for the fact that the extreme correlations in the original sample are likely overstated and more likely to occur in resamples than in (true) new samples from the population. Nonetheless, the extent of the improvement depends on the value of the randomization parameter, c. In particular, we found that although performance in full AUC was relatively insensitive to the choice of c (a result consistent with Meinshausen and Bühlmann [2010], who advocated choosing $c \in [0.2, 0.8]$), performance in the initial AUC, our preferred metric, was far more sensitive to c: Randomized penalization improved on constant penalization only for $c \in [0.6, 0.8]$. How these values generalize will require further examination, but for now we conservatively recommend manual calibration via simulation to a given SNP dataset. Moreover, to average over the additional layer of randomization, LLARRMA using randomized (as opposed to constant) penalization requires a larger number resamples to obtain comparably stable RMIP or FReMIP estimates.

Our LLARRMA-dawg procedure can be though of as a multi-SNP association model embedded within a resample model averaging scheme; as such, it can be extended in a number of ways. The association model described here uses linear regression for a quantitative phenotype, but this is trivially extended to logistic regression for case/control phenotypes (as in the original LLARRMA of Valdar et al., 2012), or to any generalized linear model to which the LASSO or group LASSO can be applied. Within our association model, we penalize effects using the group LASSO, but this could be replaced by more sophisticated type of penalization. For example, Yang et al. [2010] describe a related SNP selection method that incorporates not only additive and dominance effects but also epistasis, all via the adaptive group LASSO; their procedure uses a different parameterization of additive and dominance than ours and returns only a single best model rather than an ensemble of models, but it could in theory be embedded in our FReMA framework to yield comparable model-averaged results. Our association model focuses on grouped effects for single SNPs but it could be extended to consider different types of grouped effects, such as differential effects of local haplotype combinations, structural variants, or combinations of rare variants within a gene or LD block. For the last of these, our FReMA framework could also be used to provide model-averaged inference for existing single-solution rare variant selection procedures such as those of Zhou et al. [2010], Ayers et al. [2011], Ayers and Cordell [2013], or Larson and Schaid [2014]. In this case, a given procedure would be applied in full to multiple independent fractional resamples, with each such resample defined through a reweighting of the likelihood component (as in Equation (2)); the results of each application would then be aggregated in the manner described for Equation (7).

In summary, we describe a frequentist procedure, and its variations, to reprioritize genetic associations at loci containing multiple additive and dominance signals. The authors will provide an implementation of the proposed procedures in the R-package R/FReMA as soon as is practicable.

Description of Supplementary Material

Supplementary material contains further information about the calibration of the randomized LASSO, its efficient implementation, and simulations comparing performance of LLARRMA-dawg under alternate dominance parameterizations.

Acknowledgments

The authors thank Karen Mohlke, Yun Li, and Leslie Lange for helpful discussions. Research reported in this manuscript was supported by the following: National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award number R01GM104125 (partial support for J.S. and W.V.), National Institute of Mental Health of the NIH under award number R01MH101819 (A.B.N.), National Science Foundation (NSF) under grant number DMS-1310002 (A.B.N.), and University of North Carolina Lineberger Comprehensive Cancer Center (partial support for J.S. and W.V.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

References

Antoniou AC, Sinilnikova OM, McGuffog L, Healey S, Nevanlinna H, Heikkinen T, Simard J, Spurdle AB, Beesley J, Chen X and others. 2009. Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers. *Hum Mol Genet* 18:4442–4456.

Ayers KL, Cordell HJ. 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol* 34:879–891.

Ayers KL, Cordell HJ. 2013. Identification of grouped rare and common variants via penalized logistic regression. *Genet Epidemiol* 37:592–602.

Ayers KL, Mamasoula C, Cordell HJ. 2011. Penalized-regression-based multimarker genotype analysis of Genetic Analysis Workshop 17 data. *BMC Proc* 5 Suppl 9:S92.

Ballard DH, Cho J, Zhao H. 2010. Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet Epidemiol* 34:201–212.

Breiman L. 1996. Bagging predictors. *Machine Learning* 24:123–140.

Buckland S, Burnham K, Augustin N. 1997. Model selection: an integral part of inference. *Biometrics* 53:603–618.

Bühlmann, P, Geer S van de. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Heidelberg: Springer.

Chapman J, Clayton D. 2007. One degree of freedom for dominance in indirect association studies. *Genet Epidemiol* 271:261–271.

Clyde M, Lee H. 2001. Bagging and the Bayesian bootstrap. In: Richardson TS, Jaakkola T, editors. *Artificial Intelligence and Statistics* 2001: Proceedings of the Eighth International Workshop. Key West, FL: Morgan Kaufmann, pp. 169–174.

Dastani Z, Hivert MF, Timpson N, Perry JRB, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikäinen LP and others. 2012. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet* 8:e1002607.

Edwards SL, Beesley J, French JD, Dunning AM. 2013. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* 93:779–797.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.

Guan Y, Stephens M. 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat* 5:1780–1815.

Kim S, Morris NJ, Won S, Elston RC. 2010. Single-marker and two-marker association tests for unphased case-control genotype data, with a power comparison. *Genet Epidemiol* 34:67–77.

Kooperberg C, LeBlanc M, Obenchain V. 2010. Risk prediction using genome-wide association studies. *Genet Epidemiol* 34:643–652.

Krzanowski WJ, Hand DJ. 2009. ROC Curves for Continuous Data. Boca Raton, FL: Chapman & Hall/CRC. 1st edition.

Larson NB, Schaid DJ. 2014. Regularized rare variant enrichment analysis for case-control exome sequencing data. *Genet Epidemiol* 38:104–113.

Li Q, Zheng G, Liang X, Yu K. 2009. Robust tests for single-marker analysis in case-control genetic association studies. *Ann Hum Genet* 73:245–252.

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

McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko To, Corre T, Davies G, Kaakinen M, Lyytikäinen LP and others. 2012. Evidence of inbreeding depression on human height. *PLoS Genet* 8:e1002655.

Meier L. 2009. grplasso: fitting user specified models with Group Lasso penalty. *R package version* 0.4-2.

Meinshausen N, Bühlmann P. 2010. Stability selection. *J Roy Stat Soc B* 72:417–473.

Neale B, Ferreira M, Medland S, Posthuma D. 2008. *Statistical genetics: gene mapping through linkage and association.* New York, NY: Taylor & Francis.

Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J Roy Stat Soc B* 56:3–48.

Ober C, Abney M, McPeek MS. 2001. The genetic dissection of complex traits in a founder population. *Am J Hum Genet* 69:1068–1079.

Okamoto K, Tokunaga K, Doi K, Fujita T, Suzuki H, Katoh T, Watanabe T, Nishida N, Mabuchi A, Takahashi A and others. 2011. Common variation in GPC5 is associated with acquired nephrotic syndrome. *Nat Genet* 43:459–463.

Org E, Eyheramendy S, Juhanson P, Gieger C, Lichtner P, Klopp N, Veldre G, Döring A, Viigimaa M, Sõber S and others. 2009. Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Hum Mol Genet* 18:2288–2296.

Purcell S, Neale B, Toddbrown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Debakker P, Daly M. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.

R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rubin DB. 1981. The Bayesian bootstrap. *Ann Stat* 9:130–134.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78:629–644.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 44:825–830.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* 3:1296–1308.

Stephens M, Balding DJ. 2009. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10:681–690.

Strange A, Capon F, Spencer CCA, Knight J, Weale ME, Allen MH, Barton A, Band G, Bellenguez C, Bergboer JG and others. 2010. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat Genet* 42:985–990.

Su Z, Marchini J, Donnelly P. 2011. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 27:1–2.

The International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.

Utz H, Melchinger A, Schön, C. 2000. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* 154:1839–1849.

Valdar W, Holmes CC, Mott R, Flint J. 2009. Mapping in structured populations by resample model averaging. *Genetics* 182:1263–1277.

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J. 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* 38:879–887.

Valdar W, Sabourin J, Nobel A, Holmes CC. 2012. Reprioritizing genetic associations in hit regions using lasso-based resample model averaging. *Genet Epidemiol* 36:451–462.

Wang X, Morris NJ, Schaid DJ, Elston RC. 2012. Power of single- vs. multi-marker tests of association. *Genet Epidemiol* 36:480–487.

Warren LL, Li L, Nelson MR, Ehm MG, Shen J, Fraser DJ, Aponte JL, Nangle KL, Slater AJ, Woollard PM and others. 2012. Deep resequencing unveils genetic architecture of ADIPOQ and identifies a novel low-frequency variant strongly associated with adiponectin variation. *Diabetes* 61:1297–1301.

WTCCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678.

Yang C, Wan X, Yang Q, Xue H, Yu W. 2010. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC Bioinformatics 11 Suppl* 1:S18.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N and others. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 39:645–649.

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J Roy Stat Soc B* 68:49–67.

Zheng J, Li Y, Abecasis GR, Scheet P. 2011. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol* 35:102–110.

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26:2375–2382.