5-2013

Optimal Methods for Using Posterior Probabilities in Association Testing

Keli Liu
Harvard University

Alexander Luedtke
University of California - Berkeley

Nathan L. Tintle
Dordt College, nathan.tintle@dordt.edu

Recommended Citation

Liu K, Luedtke A, and Tintle NL (2013) "Optimal methods for using posterior probabilities in association testing" Human Heredity. 75(1): 2-11. doi:10.1159/000349974

Optimal Methods for Using Posterior Probabilities in Association Testing

Abstract

Objective: The use of haplotypes to impute the genotypes of unmeasured single nucleotide variants continues to rise in popularity. Simulation results suggest that the use of the dosage as a one-dimensional summary statistic of imputation posterior probabilities may be optimal both in terms of statistical power and computational efficiency; however, little theoretical understanding is available to explain and unify these simulation results. In our analysis, we provide a theoretical foundation for the use of the dosage as a one-dimensional summary statistic of genotype posterior probabilities from any technology. *Methods:* We analytically evaluate the dosage, mode and the more general set of all one-dimensional summary statistics of two-dimensional (three posterior probabilities that must sum to 1) genotype posterior probability vectors. *Results:* We prove that the dosage is an optimal one-dimensional summary statistic under a typical linear disease model and is robust to violations of this model. Simulation results confirm our theoretical findings. *Conclusions:* Our analysis provides a strong theoretical basis for the use of the dosage as a one-dimensional summary statistic of genotype posterior probability vectors in related tests of genetic association across a wide variety of genetic disease models.

Keywords

imputation, dosage, genome-wide association studies

Disciplines

Bioinformatics | Genetics and Genomics | Statistics and Probability

Comments

This is a pre-publication author manuscript of the following final, published article: Liu K, Luedtke A, and Tintle NL (2013) “Optimal methods for using posterior probabilities in association testing” Human Heredity. 75(1): 2-11. doi:10.1159/000349974

The definitive version is published by Karger and available at http://www.karger.com/?DOI=10.1159/000349974

Title: Optimal methods for using posterior probabilities in association testing

Authors: Keli Liu[1], Alexander Luedtke[2], Nathan Tintle[3+]

1. Department of Statistics, Harvard University, Cambridge, MA, USA
2. Division of Biostatistics, University of California-Berkeley, Berkeley, CA, USA
3. Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA, USA

+Corresponding author

Full address of corresponding author:

Dr. Nathan Tintle

498 4^{h} Ave NE

Sioux Center, IA 51250

Phone: 712-722-4863

Email: nathan.tintle@dordt.edu

Short title: Optimal methods for using posterior probabilities

Key words: Imputation, dosage, genome-wide association studies

Abstract

Objective: The use of haplotypes to impute the genotypes of unmeasured single nucleotide variants continues to rise in popularity. Simulation results suggest that the use of the dosage as a one-dimensional summary statistic of imputation posterior probabilities may be optimal both in terms of statistical power and computational efficiency, however little theoretical understanding is available to explain and unify these simulation results. In our analysis, we provide a theoretical foundation for the use of the dosage as a one-dimensional summary statistic of genotype posterior probabilities from any technology.

Methods: We analytically evaluate the dosage, mode and the more general set of all one-dimensional summary statistics of two-dimensional (three posterior probabilities that must sum to 1) genotype posterior probability vectors.

Results: We prove that the dosage is an optimal one-dimensional summary statistic under a typical linear disease model and is robust to violations of this model. Simulation results confirm our theoretical findings.

Conclusions: Our analysis provides a strong theoretical basis for the use of the dosage as a one-dimensional summary statistic of genotype posterior probability vectors in related tests of genetic association across a wide variety of genetic disease models.

Introduction

Access to high-throughput genotype data has facilitated the process of identifying the genetic component of complex diseases through genome-wide association studies (GWAS). However, directly measured genotype data often only covers a fraction of known single nucleotide polymorphisms (SNPs). Increasingly, genetic analyses leverage linkage disequilibrium (LD) to impute untyped SNPs. Analysis at untyped SNPs using LD information from reference panels, such as The International HapMap Project (T. I. H. 3 Consortium, 2010) or the 1000 genomes project (T. 1000 G. P. Consortium, 2010), has already yielded promising disease loci for major depressive disorder (Sullivan, Patrick F, de Geuss, Eco JC, Willemsen & James, Michael R, Smit, Jan H, Zandbelt, Tim, Arolt, Volker, Baune, 2009), Crohn's disease (Barrett et al., 2008), and prostate cancer (Zabaleta et al., 2009) among others.

While many imputation algorithms exist (e.g., MaCH (Li, Yun, Willer, Cristen J, Ding, Jun, Scheet, Paul, Abecasis, 2010), IMPUTE (Howie, Donnelly, & Marchini, 2009), among others), most algorithms generate a set of three posterior probabilities for each individual at each imputed SNP, representing the relative likelihood that the individual is actually each of the three possible genotypes at the SNP locus. While some exceptions exist (e.g., Lin et al. 2008), Hu and Lin (2010)), most researchers attempt to use a one-dimensional summary statistic of the two-dimensional posterior probabilities vector (three posterior probabilities that must sum to 1) in place of the (unknown) true genotype in downstream statistical analyses. Common choices for the one-dimensional summary statistics are the mode and the weighted mean (dosage) of the three posterior probabilities. While posterior probabilities are common for imputed genotypes they also occur when using next-generation sequencing data and SNP microarray technology.

Recently, extensive simulations demonstrated that the dosage retained enough information that, in most realistic settings, the use of computationally intensive mixture models which account for the entire posterior probability vector improved power negligibly over a simpler, faster analysis using the dosage (Zheng, Li, Abecasis, & Scheet, 2011). Furthermore, the dosage consistently outperformed the mode.

Despite these simulation results, little theoretical work has been conducted to consider whether the dosage will always perform optimally relative to the mode. Furthermore, while the dosage is a reasonable choice of one-dimensional summary statistic, it is unknown if more optimal summary statistics are available. In the following manuscript we provide analytic proof that across a variety of disease models dosage will always outperform the mode. We then show that the dosage is equivalent to the optimal one-dimensional summary statistic up to a perturbance term which is essentially zero in all practical situations. We confirm our analytic results using simulation.

Methods

The following subsections are organized as follows. First, we provide an overview of our notation and the main genetic disease model under consideration. We then provide proof that the score test using the dosage is equivalent to the score test using the entire vector of posterior probabilities. We then show that the dosage outperforms the mode and is, in fact, the optimal choice of one-dimensional summary statistic across linear genetic disease models, and is a robust choice for non-linear models. We conclude the methods section by describing simulation analyses used to confirm the analytic results.

Basic notation and disease model

Individual genotypes provided by imputation software, as well as some SNP microarray and sequencing technology, are typically provided for each individual, i, as a vector of three posterior probabilities, $\alpha_i \triangleq (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})$, where α_{ik} is the posterior probability that individual i has k minor alleles, $k = 0,1,2$ at a SNP of interest. The vector of posterior probabilities, α_i, can be interpreted as suggesting that the true minor allele count for individual i, denoted x_i, can be modeled as being a single random draw from a multinomial distribution with probabilities indicated by α_i. We assume that α_i is known for each individual. Let y_i be an indicator for disease status and let the probability of disease for individual i be given as $\pi_i(x_i)$. A general formulation for the disease-genotype relationship is $y_i|x_i \sim Bern(\pi_i(x_i))$ where $\pi_i = f(x_i, \beta)$. We assume that f is a smooth function and β is a parameter vector constrained so that the range of f is some subset of the unit interval. We will use the term additive model, to mean that the function f depends on x_i and β only through a term of the form $\beta_0 + \beta_1 x_i$. In this manuscript we explore two types of additive models (1) A linear additive model: $\pi_i = \beta_0 + \beta_1 x_i$ (2) and a (nonlinear) logistic additive model, $\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. When x_i is unobserved, it is common to naively plug in a one-dimensional summary statistic such as the mode, $m_i \triangleq \arg\max_k \alpha_{ik}$, or the dosage, $d_i \triangleq \alpha_{i1} + 2\alpha_{i2}$, for the genotype in the disease model.

Score Test Using the Posterior Probability Vector

Intuitively, we anticipate a loss of information when summarizing the entire posterior probability vector with a one-dimensional summary statistic. However, we will now show for the commonly used logistic model that a score test using the dosage is equivalent to a score test from a model that incorporates the entire posterior probability vector.

The observed data likelihood of $\tilde{\pi}_i$ for a random sample of *n* individuals is given by $L(\tilde{\pi}_i) = \prod_{i=1}^{n} \tilde{\pi}_i^{y_i}(1-\tilde{\pi}_i)^{1-y_i}$, where $\tilde{\pi}_i = \sum_{k=0}^{2} \frac{\exp(\beta_0+\beta_1 k+\boldsymbol{\gamma}'\boldsymbol{z_i})}{1+\exp(\beta_0+\beta_1 k+\boldsymbol{\gamma}'\boldsymbol{z_i})} \alpha_{ik}$ and $\boldsymbol{z_i}$ is a vector of covariates. The mixture results because we've marginalized over the x_i. The score component, denoted $U(\beta_1)$, which is the gradient of the log-likelihood for β_1, is

$$U(\beta_1) = \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} \frac{\partial l}{\partial \tilde{\pi}_i}\frac{\partial \tilde{\pi}_i}{\partial \beta_1} = \sum_{i=1}^{n} \left(\frac{y_i}{\tilde{\pi}_i} - \frac{1-y_i}{1-\tilde{\pi}_i}\right)\left(\sum_{k=0}^{2} \frac{\exp(\beta_0+\beta_1 k+\boldsymbol{\gamma}'\boldsymbol{z_i})}{1+\exp(\beta_0+\beta_1 k+\boldsymbol{\gamma}'\boldsymbol{z_i})} k\alpha_{ik}\right)$$

where $l(\cdot)$ is the log-likelihood. Thus, under the null hypothesis of no association, $\beta_1 = 0$ and $\tilde{\pi}_i = \pi_0$ for $\tilde{\pi}_{i0} = \frac{\exp(\beta_0+\boldsymbol{\gamma}'\boldsymbol{z_i})}{1+\exp(\beta_0+\boldsymbol{\gamma}'\boldsymbol{z_i})}$. The score component is:

$$U(\beta_1 = 0) = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_0} - \frac{1-y_i}{1-\pi_0}\right)\left(\alpha_{i1}\pi_0(1-\pi_0) + 2\alpha_{i2}\pi_0(1-\pi_0)\right)$$

$$= \sum_{i=1}^{n} (y_i(1-\pi_0) - (1-y_i)\pi_0)(\alpha_{i1} + 2\alpha_{i2}) = \sum_{i=1}^{n} (y_i - \pi_0)d_i$$

Next we consider the score component for the model which substitutes the dosage for the unknown genotype. The model utilizing the dosage can be written $\pi_i = \frac{\exp(\beta_0+\beta_1 d_i+\boldsymbol{\gamma}'\boldsymbol{z_i})}{1+\exp(\beta_0+\beta_1 d_i+\boldsymbol{\gamma}'\boldsymbol{z_i})}$. In this case, the score component is

$$U(\beta_1) = \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} \frac{\partial l}{\partial \tilde{\pi}_i}\frac{\partial \tilde{\pi}_i}{\partial \beta_1} = \sum_{i=1}^{n} \left(\frac{y_i}{\tilde{\pi}_i} - \frac{1-y_i}{1-\tilde{\pi}_i}\right)\left(\frac{\exp(\beta_0+\beta_1 d_i)}{(1+\exp(\beta_0+\beta_1 d_i))^2} d_i\right).$$

And thus, $U(\beta_1 = 0) = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_{i0}} - \frac{1-y_i}{1-\pi_{i0}}\right)(\pi_{i0}(1-\pi_{i0})d_i) = \sum_{i=1}^{n} (y_i - \pi_{i0})d_i$ where again $\tilde{\pi}_{i0} = \frac{\exp(\beta_0+\boldsymbol{\gamma}'\boldsymbol{z_i})}{1+\exp(\beta_0+\boldsymbol{\gamma}'\boldsymbol{z_i})}$.

We note that a score test is the score component, *U*, divided by the negative derivative of the score component, *I*, which is known as the observed information. Thus, because *U* is equivalent and the null models under either formulation admit the same distribution for *U*, it

follows that the score tests for association are equivalent. When covariates are present, U and I depend on $\boldsymbol{\gamma}$. The score statistic is computed by substituting $\widehat{\boldsymbol{\gamma}}_{\mathbf{0}}$, the MLE estimated under the null model, in place of $\boldsymbol{\gamma}$.

Optimal choice of a one dimensional summary statistic

While we have demonstrated that the score test using the dosage is equivalent to a score test using the entire vector of posterior probabilities, we have not considered the power of such an approach. The score test is known to perform well asymptotically, but we have no reason to assume that the dosage will perform well in finite samples. In the following sections we explore intuitive choices for $\boldsymbol{g}$, a general one-dimensional summary of the posterior probabilities, and then derive its optimal value.

Explicit expression for the non-centrality parameter

In *Appendix A*, we assume a logistic model and provide the score statistic for any one-dimensional summary, $\boldsymbol{g}$, of $\boldsymbol{x}$, the true genotype. We derive the statistic under the logistic model because this is the most popular model choice in applied work. All subsequent power calculations are performed with respect to this class of test statistics. Our goal is to evaluate the power across a range of true models, which may or may not be logistic, seeing how the optimal summary depends on various assumptions about the true model. An expression is given for the noncentrality parameter when the true model has a general form, $\pi_i = f(x_i, \beta)$, and an analytically useful version results when we assume the true model is linear. Since a larger noncentrality parameter is a necessary and sufficient condition for a more powerful test, in the

proceeding sections, we will use the noncentrality parameter to compare the efficiency of various one-dimensional summaries, $\boldsymbol{g}$.

Dosage Beats Mode

In *Appendix B*, we prove that the dosage is more highly correlated with the true genotype than the mode. This result makes no assumptions about the form of the true disease model (e.g., linear or logistic) and holds for all finite sample sizes. We note, however, that a higher correlation between the imputed genotype and the true genotype does not automatically imply a more powerful test. To draw conclusions about higher power from higher correlation alone, one must also assume that the true disease model is linear: $\pi_i = \beta_0 + \beta_1 x_i$. We consider non-linear models more explicitly later (*Nonlinear Disease Models*).

Optimal Summary of Posterior Distribution

Although the superiority of the dosage to the mode is an important result, we have not yet demonstrated that the dosage is an optimal one-dimensional summary statistic. In *Appendix C*, we show that the score test which results from using the dosage is essentially identical to a score test using the optimal one-dimensional statistic. In *Appendix C*, we start by noting that the optimal one-dimensional statistic yields the test with the largest noncentrality parameter. Finding the optimal statistic is therefore equivalent to finding the statistic which maximizes the noncentrality parameter defined in *Appendix A*. Results from perturbation theory show that the dosage is nearly identical to the optimal statistic in all realistic situations. When the true disease model is linear, it follows that the score test using the dosage is essentially optimal.

Taylor series expansions show that an additive logistic disease model is well approximated by a linear model when disease prevalence is low for all genotype groups or the SNP effect size is small. Additionally, we expect the projection of the additive logistic model onto the space of all linear models to result in a model respecting the range of π_i, namely some subset of the unit interval, since x_i is bounded and only takes three values. For cases in which this linear approximation is inadequate, our results in the section *Score Test Using the Posterior Probability Vector* still apply. These results serve as insurance that the dosage will perform strongly in the case of an additive logistic model, that it will be asymptotically optimal since the score test is asymptotically most powerful. However, there is no longer any claim on optimality in finite samples when the additive logistic model is not well approximated by a linear model.

Nonlinear Disease Models and Covariates

The previous sections assumed either a linear or approximately linear disease model. While the assumption of approximate linearity is common in practice with the use of a logistic model, we now consider non-linear modes of inheritance. We first analyze the situation without covariates. In *Appendix D* we derive the optimal score test for this case. As in the case of a linear model, the optimal statistic for a nonlinear model is well approximated by a simple one-dimensional statistic. Specifically, an approximation of the optimal statistic is given by a linear combination of the dosage and a generalization of the dosage which we call the second-order dosage. This linear combination is given by $d_i^{(1)} + \xi d_i^{(2)}$, where $d_i^{(j)} = \alpha_{i1} + 2^j \alpha_{i2}$ represents the j^{th} order dosage (which governs the j^{th} order effect) for $j = 1,2$ and ξ represents the ratio of the second to first order effects, a measure of non-linearity.

To build some intuition about ξ, let us examine two common non-linear models. For a dominant disease model use $\alpha_{1i} + \alpha_{2i}$ instead of the dosage, and for a recessive disease model use α_{2i} instead of the usual dosage. For more complex models, we can formulate the intuition behind ξ as follows. The posterior distribution of the genotype can be indexed by two parameters: the mean and the variance. For linear and approximately linear models, one can simply pretend that $x_i = d_i$ with little cost. For nonlinear models, the cost is non-trivial. The variance in the posterior distribution should then inform us on an individual basis of the cost of the assumption $x_i = d_i$, and allow us to adjust the weight of evidence accordingly.

Appendix D shows that the only change to our analysis through the inclusion of covariates is to allow the nonlinearity of the SNP effect to depend on the values of the covariates. Implementation wise, the optimal linear combination is now $d_i^{(1)} + \xi(z_i) d_i^{(2)}$ where the measure of effect nonlinearity, ξ, depends on z_i. Such complications can be avoided if the SNP and covariate effects are additively separable, $\pi_i = g(x_i) + h(\boldsymbol{z}_i)$, in which case a common ξ suffices to summarize nonlinearity of the SNP effect.

Continuous Traits

In *Appendix E*, we derive the optimal score test for normally distributed continuous traits which are linear functions of the genotype and a set of covariates. Thus we have analytically shown that the dosage is the optimal statistic for the additive continuous traits model considered in Zheng et al. (2011).

Simulation

To verify the theoretical results empirically, we calculated power using simulated data. We considered three different characteristics of SNPs: (1) the r^2 coefficient between the dosages and the genotype (Note: this r^2 coefficient is the value that MACH approximates with its r^2 imputation quality measure (Li, Yun, Willer, Cristen J, Ding, Jun, Scheet, Paul, Abecasis, 2010)) (2) the minor allele frequency, which unless otherwise stated was set at 0.1 and (3) the odds ratio under an additive disease model. We consider values of r^2ranging from 0.1 to 1, MAF ranging from 0.05 to 0.5 and odds ratios ranging from 1.0 to 2.0. For each simulation setting 10,000 SNPs were simulated with 1000 cases and 1000 controls. Disease prevalence was fixed at 50% among individuals with no risk alleles. Unless otherwise stated, power was calculated at the 5% significance level using the asymptotic distributions of the score tests.

For each SNP and each individual i we compute posterior probabilities α_i by sampling from a Dirichlet distribution, where $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(\frac{(1-q)^2}{c}, \frac{2q(1-q)}{c}, \frac{q^2}{c})$. We let q indicate the minor allele frequency of the SNP and c be a nonnegative constant chosen to obtain the desired r^2. Varying c does not appear to greatly modify the standard deviation of the r^2 coefficient, which ranges between 0.02 and 0.03 across all simulations. An individual's genotype x_i was determined by sampling from a multinomial distribution with probabilities indicated by the vector of posterior probabilities α_i.

Results

We conducted simulation analyses to confirm the theoretical findings described earlier. In the following sections we briefly describe the results of these simulation analyses. Figure 1 empirically demonstrates that the score test for the dosage is uniformly more powerful than the test for mode. In this setting we note that the power for the true genotype test is much higher than

the power for both the dosage and mode score tests. However, this is not surprising given the relatively low imputation quality (r^2=0.6), used in this graph.

As expected, as imputation quality increases, power increases (Figure 2). Furthermore, the power of the dosage and mode tests approaches the power of the linear trend test using the true genotype as r^2 increases. We note that, for $r^2 = 1$, all three tests are identical and thus obtain the same power. On the other hand, for low r^2 the dosage and mode contain little information about the true genotype, and so low power is obtained. Nonetheless, it is interesting to see that the dosage outperforms the mode even in this setting.

Figure 3 shows that the power of all methods increases as the log odds ratio increases. When the odds ratio is 1 the power is equivalent to type I error rate. Importantly, all methods control type I error with empirical type I error rates equal to the nominal rate of 0.05 (detailed results not shown). For larger odds ratios we find the expected result that true genotypes are more powerful than dosages, which are in turn more powerful than the mode. As odds ratios grow sufficiently large all methods have power approaching 1, though the power rankings of the three methods remains.

Lastly, Figure 4 shows that power decreases as minor allele frequency decreases. While all tests of association are low powered at a minor allele frequency of 0.05, relative power ordering still holds for the methods, with the true genotype yielding the highest power at 0.40, the dosage yielding the second highest power at 0.26, and the mode yielding the lowest power at 0.22.

We also compared the dosage to the optimal summary statistic given as a dominant eigenvector in *Appendix C*. Simulations showed that the power and type I error were virtually identical using the two statistics (detailed results not shown).

Discussion

Previous work has shown that the computational overhead may not be worth the modest power gain from using the entire vector of posterior probabilities instead of the weighted mean posterior probability (dosage). In our analysis, we provided analytic proof that the dosage is essentially equivalent to the optimal choice of a single summary statistic in all practical situations across a range of genetic disease models, far exceeding the power obtained from using the modal posterior probability. These results were confirmed via simulation.

There are a number of important implications of these conclusions. First, while theoretical results and simulations considered the score test, due to the asymptotic equivalence, the optimality of the dosage extends to the related likelihood ratio and Wald tests. Furthermore, as considered in *Appendix D*, extensions to models including covariates show similar results, unless the effects of covariates are very large.

Superficially, our results may seem to depend (i) on the assumption that the true model is linear or approximately linear and (ii) on asymptotic approximations via Taylor expansions. However, two facts show that our results should have broad applicability across a range of models and for finite samples. First, the dosage is the one dimensional summary most highly correlated with the true genotype. This holds for all sample sizes and regardless of the true disease model. Second, naïvely assuming $\pi_i = \frac{\exp(\beta_0+\beta_1 d_i)}{1+\exp(\beta_0+\beta_1 d_i)}$ in place of the actual mixture model $\pi_i = \sum_{k=0}^{2} \frac{\exp(\beta_0+\beta_1 k)}{1+\exp(\beta_0+\beta_1 k)} \alpha_{ik}$ does not change the resulting score statistic. Since it is well known that an additive logistic disease model is robust to misspecification (Friedlin B et al.,

2002), this suggests that the score statistic from the naïve model, $\pi_i = f(d_i)$, well approximates the score statistic from the true mixture model, $\pi_i = \sum_{i=0}^{2} f(k)$.

Additionally, as shown in *Appendix D*, in many realistic non-linear models, the dosage remains a nearly optimal choice of one-dimensional statistic because the degree of nonlinearity is negligible. Even when non-linearity of the SNP effect is appreciable, *Appendix D* shows that to the extent that the non-linearity can be explicitly described, simple adjustments to the optimal one-dimensional summary can be made to preserve efficiency. Trivially, for a dominant disease model use $\alpha_{1i} + \alpha_{2i}$ instead of the dosage, and for a recessive disease model use α_{2i} instead of the usual dosage. A method for choosing the optimal summary statistic for more complex models can be found in *Appendix D*. In additional simulation analyses, not shown here, the dosage showed robustness across a wide-variety of non-linear models, with robustness related to the extent of non-linearity.

For continuous responses, Zheng et al. (2011) implemented a mixture model utilizing the entire vector of posterior probabilities. In analyses not reported here, we did similarly with an E-M algorithm and found that, like Zheng, dosage performed similarly with far less computational expense. In *Appendix E* we outline the proof of optimality of the dosage under a normal linear model for a quantitative response. While we did not provide simulations here, extensive consideration of this model and deviations from it can be found in Zheng et al. (2011). These extensive simulations also consider cases with sample sizes as low as 50 (Zheng et al.; Figure 3) suggesting robustness to the asymptotic assumptions underlying portions of the results shown here.

While we do not consider the new class of rare variant tests explicitly, our results may be extendable to two classes of rare variant tests, with a word of caution. For rare variant tests

which collapse rare variants into a single "super variant", the dosage is given by the probability that any one of the included variant sites contains the rare variant, that is by $1 - \prod_j \alpha_{0ij}$, where α_{0ij} represents the posterior probability for person i not having a rare variant at site j. For rare variant tests which regress the total number of rare variants present across a set of variant sites, the modified dosage is given by $\sum_j d_{ij}$. The former is really a special case of the latter based on the approximation $\sum_j x_{ij} \approx 1\{\sum_j x_{ij} > 1\}$. Suppose that the variants enter into the disease model additively, $\sum_j \beta_j x_{ij} = \beta \sum_j \frac{\beta_j}{\beta} x_{ij}$. Then our results apply by thinking about the imputation of $\sum_j \frac{\beta_j}{\beta} x_{ij}$ which now depends not only on the posterior probabilities for each x_{ij} but also on the nuisance parameters $\frac{\beta_j}{\beta}$. How to effectively estimate the nuisance parameters remains an area of active research. However, given the nuisance parameters, our results suggest that a near optimal summary would be $\sum_j \frac{\beta_j}{\beta} d_{ij}$. In particular, one may justify use of $\sum_j d_{ij}$ by assuming homogeneity of effects across variants. Caution needs to be taken for small sets of variant sets or sample sizes, however, because in these cases the perturbation term by which the dosage differs from the optimal summary may be nontrivial. Simulation studies and further analysis of these rare variant strategies, along with consideration of the recently proposed class of variance-components tests, is needed.

Few assumptions are required on the posterior probabilities in order for the results described here to be valid. In particular, posterior probabilities, while commonly obtained from imputation, can also be obtained from both SNP microarray and next-generation sequencing technologies. The analytic calculations shown here directly extend to these platforms. The main necessary assumption about the posterior probabilities is that they are correctly calibrated—namely, that the vector of posterior probabilities, α_i, can be interpreted as suggesting that the

true minor allele count for individual i, denoted x_i, can be modeled as being a single random draw from a multinomial distribution with probabilities indicated by α_i. While this interpretation is almost uniformly made in practice, any systematic technological bias could impact this interpretation, making the analytic conclusions provided above no longer hold.

The dosage is commonly used as a shortcut to use of a wide-class of statistical methods, which assume knowledge of the true genotypes. We provide analytic justification of its use across a wide variety of genetic models.

Acknowledgements

We acknowledge the work of Jennifer James and Nathaniel Bowerman in early phases of this project. This work was funded by the National Human Genome Research Institute (R15HG004543; R15HG006915). We acknowledge the use of the Hope College parallel computing cluster for assistance in data simulation and analysis.

References

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics*, *40*(8), 955–62. doi:10.1038/ng.175

Consortium, T. 1000 G. P. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–73. doi:10.1038/nature09534

Consortium, T. I. H. 3. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–8. doi:10.1038/nature09298

Friedlin B., Zheng, G., Li, Z., & Gastwirth, J.L. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity*, 53, 146-52.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, *5*(6), e1000529. doi:10.1371/journal.pgen.1000529

Hu, Y. J., & Lin, D. Y. (2010). Analysis of untyped SNPs: maximum likelihood and imputation methods. *Genetic epidemiology*, *34*(8), 803–15. doi:10.1002/gepi.20527

Lang, B. (2000). Direct Solvers for Symmetric Eigenvalue Problems. In J. Grotendorst (Ed.), *Modern Methods and Algorithms of Quantum Chemistry* (pp. 231–259). Julich: John von Neumann Institute for Computing.

Li, Yun, Willer, Cristen J, Ding, Jun, Scheet, Paul, Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to esimate haplotypes and unobserved genotypes. *Genetic epidemiology*, *34*(8), 816–834. doi:10.1002/gepi.20533.MaCH

Lin, D. Y., Hu, Y., & Huang, B. E. (2008). Simple and Efficient Analysis of Disease Association with Missing Genotype Data. *American Journal of Human Genetics*, *82*(February), 444–452. doi:10.1016/j.ajhg.2007.11.004.

Sullivan, Patrick F, de Geuss, Eco JC, Willemsen, G., & James, Michael R, Smit, Jan H, Zandbelt, Tim, Arolt, Volker, Baune, B. T. (2009). Genomewide association for major depressive disorder: A possible role for the presynaptic protein Piccolo. *Molecular Psychiatry*, *14*(4), 359–375. doi:10.1038/mp.2008.125.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequence data with the sequence kernel association test. *American Journal of Human Genetics 89*(July), 82-93. doi: 10.1016/j.ajhg.2011.05.029

Zabaleta, J., Su, L. J., Lin, H.-Y., Sierra, R. a, Hall, M. C., Sartor, a O., Clark, P. E., et al. (2009). Cytokine genetic polymorphisms and prostate cancer aggressiveness. *Carcinogenesis*, *30*(8), 1358–62. doi:10.1093/carcin/bgp124

Zheng, J., Li, Y., Abecasis, G. R., & Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic epidemiology*, *35*(2), 102–10. doi:10.1002/gepi.20552

Appendix A

Much of the derivation of the non-centrality parameter follows Wu et al. (2011). The logistic model utilizing a general one-dimensional summary of the posterior genotype distribution, g_i, can be written as $y_i|g_i \sim Bern\big(\pi_i(g_i)\big)$, where $logit(\pi_i) = \beta_0 + \beta_1 g_i$. Following arguments made in the main text (*Methods: Score Test Using the Posterior Probability Vector*), we can denote the squared score component as $(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}})'\boldsymbol{g}\boldsymbol{g}'(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}})$ where $\boldsymbol{y} = (y_1, \dots, y_n)'$, $\boldsymbol{g} = (g_1, \dots, g_n)'$ and $\widehat{\boldsymbol{\pi}}_0$ as the vector of disease probabilities estimated under the null hypothesis. We note that when the SNP is typed, $g_i = x_i$ and the above corresponds to the Armitage linear trend test.

Define $\boldsymbol{\mu_\beta} \triangleq (\pi_1 - \pi_0, \dots, \pi_n - \pi_0)'$ as the vector of differences between the true disease probability and the null disease probability. Under H_0, we have $\boldsymbol{\mu_\beta} = \boldsymbol{0}_n$. Let $\boldsymbol{V} \triangleq Cov(\boldsymbol{y}) = diag\big(\pi_i(1 - \pi_i)\big)$. The squared score component can be rewritten as follows

$$
\begin{aligned}
&(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_0)'\boldsymbol{g}\boldsymbol{g}'(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}}) \\
&\quad = \big(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_0 - \boldsymbol{\mu_\beta} + \boldsymbol{\mu_\beta}\big)'\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{g}\boldsymbol{g}'\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\big(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}} - \boldsymbol{\mu_\beta} + \boldsymbol{\mu_\beta}\big) \\
&\quad = \big(\boldsymbol{z} + \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\big)'\boldsymbol{K}\big(\boldsymbol{z} + \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\big)
\end{aligned}
$$

where $\boldsymbol{z} \triangleq \boldsymbol{V}^{-1/2}\big(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}} - \boldsymbol{\mu_\beta}\big)$, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} \triangleq \boldsymbol{V}^{-1/2}\boldsymbol{\mu_\beta}$, and $\boldsymbol{K} \triangleq \boldsymbol{V}^{1/2}\boldsymbol{g}\boldsymbol{g}'\boldsymbol{V}^{1/2}$. Note that z_i has mean 0 and variance 1. A spectral decomposition on $\boldsymbol{K}$ gives $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ where $\boldsymbol{U} = (\boldsymbol{u_1}, \dots, \boldsymbol{u_n})$ and $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_n)$. u_i are the orthonormal eigenvectors of $\boldsymbol{K}$ corresponding to the eigenvalues λ_i. Because $\boldsymbol{g}$ is $n \times 1$, only one of the eigenvalues is non-zero and we take this to be λ_1. Asymptotically $\boldsymbol{u_1'}\big(\boldsymbol{z} + \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\big) \sim N\big(\boldsymbol{u_1'}\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, 1\big)$. Thus,

$$
(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}})'\boldsymbol{g}\boldsymbol{g}'(\boldsymbol{y} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}}) = \big(\boldsymbol{z} + \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\big)'\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\big(\boldsymbol{z} + \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\big) \sim \boldsymbol{\lambda_1}\boldsymbol{\chi_1^2}(\boldsymbol{\delta})
$$

where δ is the noncentrality parameter, and $\delta \triangleq \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\boldsymbol{u}_{\boldsymbol{1}}\boldsymbol{u}_{\boldsymbol{1}}'\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \left(\boldsymbol{u}_{\boldsymbol{1}}'\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\right)^2$. Note that $\boldsymbol{u}_{\boldsymbol{1}} \cdot \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \cos\theta \, \|\boldsymbol{u}_{\boldsymbol{1}}\| \left\|\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\right\| = \cos\theta \left\|\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\right\|$ since $\boldsymbol{u}_{\boldsymbol{1}}$ is orthonormal and $\cos\theta \left\|\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\right\|$ is simply the length of the projection of $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ in the direction of $\boldsymbol{u}_{\boldsymbol{1}}$. In this case $u_1 = \frac{\boldsymbol{V}^{\boldsymbol{1/2}}\boldsymbol{g}}{\|\boldsymbol{V}^{\boldsymbol{1/2}}\boldsymbol{g}\|}$.

Thus the quantity which determines power is $\cos\theta$ where θ is the angle between $\boldsymbol{V}^{\boldsymbol{1/2}}\boldsymbol{g}$ and $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$, i.e. how well the summary of the posterior genotype distribution $\boldsymbol{g}$ is aligned with the ways in which the truth deviates from the null hypothesis, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}_{\boldsymbol{0}}$.

Consider the special case of a disease model which is truly linear, $\pi_i = \pi_0 + \beta_1 x_i$. Under this model $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \beta_1\boldsymbol{x}$. The noncentrality parameter is $\left(\boldsymbol{u}_{\boldsymbol{1}}'\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}\right)^2 = (\beta_1\boldsymbol{g}'\boldsymbol{x})^2 = \left(\beta_1 \cos\theta_{\boldsymbol{g},\boldsymbol{x}} \, \|\boldsymbol{g}\|\|x\|\right)^2$. Note that the key quantity is $\cos\theta_{\boldsymbol{g},\boldsymbol{x}}$ or how well aligned our summary $\boldsymbol{g}$ is with the true genotype vector $\boldsymbol{x}$.

Appendix B

We now consider the special cases of the score test from *Appendix A* where the one-dimensional summary statistic g_i is equal to d_i (dosage) or m_i (mode). $\cos\theta_{\boldsymbol{g},\boldsymbol{x}}$ is then the observed correlation between g_i and the true genotype x_i and we show that the dosage is always more highly correlated with the true genotype than is the mode. In this Appendix we assume that the posterior probability vectors $\alpha_i = (\alpha_{0i}, \alpha_{1i}, \alpha_{2i})$ are drawn i.i.d from some arbitrary distribution on the 2-simplex. We note that this implies d_i and m_i are now both random variables whereas above when α_i was treated as fixed, the dosage and mode were also fixed. In other sections of this paper we condition on α_i and thus we are able to treat them as constant.

Using the law of total covariance

$$Cov(x_i, d_i) = E\big(Cov(x_i, d_i|\alpha_i)\big) + Cov\big(E(x_i|\alpha_i), E(d_i|\alpha_i)\big) = Var(d_i)$$

Because given α_i, d_i is a constant and thus $Cov(x_i, d_i|\alpha_i) = 0$ and $d_i = E(x_i|\alpha_i)$. A second application of the law of total covariance gives:

$$Cov(x_i, m_i) = E\big(Cov(x_i, m_i|\alpha_i)\big) + Cov\big(E(x_i|\alpha_i), E(m_i|\alpha_i)\big) = Cov(d_i, m_i)$$

Let r_{mean} represent $Cor(x_i, d_i)$ and r_{mode} represent $Cor(x_i, m_i)$. Thus, by substitution, we have

$$\begin{aligned} r_{mean}^2 - r_{mode}^2 &= \frac{Cov(x_i,d_i)}{\sqrt{Var(x_i)Var(d_i)}} - \frac{Cov(x_i,m_i)}{\sqrt{Var(x_i)Var(m_i)}} \\ &= \frac{Var(d_i)\sqrt{Var(m_i)} - Cov(x_i,m_i)\sqrt{Var(d_i)}}{\sqrt{Var(x_i)Var(d_i)Var(m_i)}} \\ &= \frac{\sqrt{Var(d_i)Var(m_i)} - Cov(d_i,m_i)}{\sqrt{Var(x_i)Var(m_i)}} \\ &\geq 0 \end{aligned}$$

since $Cov(d_i, m_i) \leq \sqrt{Var(d_i)Var(m_i)}$ by the Cauchy-Schwarz inequality.

So far, no asymptotic arguments have been used. Asymptotics come into play only in linking r^2_{mean} to $\cos^2 \theta_{\boldsymbol{d},\boldsymbol{x}}$ and r^2_{mode} to $\cos^2 \theta_{\boldsymbol{m},\boldsymbol{x}}$, where r^2_{mean} and r^2_{mode} are the population counterparts of the sample quantities, $\cos^2 \theta_{\boldsymbol{d},\boldsymbol{x}}$ and $\cos^2 \theta_{\boldsymbol{m},\boldsymbol{x}}$. The justification for using dosage over mode in any particular sample depends on the inequality $\cos^2 \theta_{\boldsymbol{d},\boldsymbol{x}} \geq \cos^2 \theta_{\boldsymbol{m},\boldsymbol{x}}$ rather than the inequality $r^2_{mean} \geq r^2_{mode}$. As $n \to \infty$, $\cos^2 \theta_{\boldsymbol{d},\boldsymbol{x}} \to r^2_{mean}$ and $\cos^2 \theta_{\boldsymbol{m},\boldsymbol{x}} \to r^2_{mode}$ (Lang, 2000). Thus as $n \to \infty$, $\cos^2 \theta_{\boldsymbol{d},\boldsymbol{x}} \geq \cos^2 \theta_{\boldsymbol{m},\boldsymbol{x}}$ almost surely. And so, based on the equation for the noncentrality parameter shown in *Appendix A*, the test using the dosage has a larger noncentrality parameter than the test using the mode, implying that the score test using the dosage test has more power than the test using the mode.

Appendix C

For the set of linear trend tests using the statistic $(\boldsymbol{y}-\widehat{\boldsymbol{\pi}}_{\boldsymbol{0}})'\boldsymbol{g}\boldsymbol{g}'(\boldsymbol{y}-\widehat{\boldsymbol{\pi}}_{\boldsymbol{0}})$, we derive the optimal $\boldsymbol{g}$, where we define optimal to be the $\boldsymbol{g}$ which yields the most powerful score test.

Thus, to find the optimal $\boldsymbol{g}$ we wish to maximize the noncentrality parameter $\boldsymbol{u_1'\widetilde{\mu}_\beta\widetilde{\mu}_\beta'u_1}$ (see *Appendix A*) subject to the constraint that $\|u_1\| = 1$. Under the linear disease model, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \beta_1\boldsymbol{V}^{-1/2}\boldsymbol{x}$ where β_1 is a multiplicative constant that is irrelevant to the optimization problem. Thus we wish to solve the optimization problem given as follows:

$$\max_{u\in\mathbb{R}^n} E\left(\boldsymbol{u}'\left(\boldsymbol{V}^{-1/2}\boldsymbol{x}\right)\left(\boldsymbol{V}^{-1/2}\boldsymbol{x}\right)'\boldsymbol{u}\right) \text{ subject to } \boldsymbol{u}'\boldsymbol{u} = 1$$

Because $\boldsymbol{u} = \boldsymbol{V}^{1/2}\boldsymbol{g}$ and the optimal $\boldsymbol{g}$ is unique only up to scaling (significance of the test does not depend on how we scale $\boldsymbol{g}$), we can reformulate the problem as:

$\max_{\boldsymbol{g}\in\mathbb{R}^n} E(\boldsymbol{g}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{g})$ subject to $\boldsymbol{g}'\boldsymbol{g} = 1$. The first-order conditions are

$$(E(\boldsymbol{x})E(\boldsymbol{x})' + Cov(\boldsymbol{x}) - \lambda\boldsymbol{I})\boldsymbol{g} = 0$$

Thus $\boldsymbol{g}$ is dominant eigenvector of

$$E(\boldsymbol{x})E(\boldsymbol{x})' + Cov(\boldsymbol{x})$$

$$= \boldsymbol{d}\boldsymbol{d}' + Cov(\boldsymbol{x})$$

where $Cov(\boldsymbol{x}) = diag(\alpha_{1i} + 4\alpha_{2i} - d_i^2)$.

Let $\boldsymbol{g}^*$ be the dominant eigenvector of $\boldsymbol{d}\boldsymbol{d}'$**,** which ignores the covariance term. Below, we justify why ignoring the covariance term has negligible effect in most situations. Note that $\boldsymbol{g}^* = \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|}$ and since the scaling on $\boldsymbol{g}$ does not matter for testing purposes, this suggests we take $\boldsymbol{g} = \boldsymbol{g}^* = \boldsymbol{d}$ if the covariance term can be ignored.

To justify approximating $\boldsymbol{g}$ with $\boldsymbol{g}^*$, in essence why it is acceptable to ignore the covariance term, we can assume without loss of generality that $\max_i[\alpha_{1i} + 4\alpha_{2i} - d_i^2] = var(x_1)$. Applying a result from perturbation theory, see Equation 6 in [11], yields :

$$\frac{1}{2}\sin 2\theta_{g^*,g} \leq \frac{\max_i \alpha_{1i} + 4\alpha_{2i} - d_i^2}{\sum_i d_i^2} \leq$$

$$\frac{var(x_1)}{E(x_1)^2 + \sum_{i=2}^n E(x_i)^2} = \left(\frac{1}{CV^2(x_1)} + \sum_{i=2}^n \frac{var(x_i)}{var(x_1)}\frac{1}{CV^2(x_i)}\right)^{-1} = \left(\sum_{i=1}^n w_i \frac{1}{CV^2(x_i)}\right)^{-1}$$

where CV is the coefficient of variation where $w_i \leq 1$.

We can interpret $\frac{1}{CV^2(x_i)}$ as a measure of of the signal to noise ratio (of the imputation process) for individual *i*. Thus, $\sum_{i=1}^n w_i \frac{1}{CV^2(x_i)}$ is a weighted sum that measures the overall precision of the imputation process, where the weights, $\frac{var(x_i)}{var(x_1)}$, serve as a standardization factor. As *n* goes to infinity, the weighted sum of the precisions goes to infinity as well, i.e. there is an accumulation of genotype information across individuals so that the angle between $\boldsymbol{g}$ and $\boldsymbol{g}^*$ goes to 0. Finally, we combine the bound with the following approximation for small angles, $\theta_{\boldsymbol{g}^*,\boldsymbol{g}} \approx \sin\theta_{\boldsymbol{g}^*,\boldsymbol{g}} \approx \frac{1}{2}\sin\theta_{\boldsymbol{g}^*,\boldsymbol{g}} \approx 0$ to conclude that $\boldsymbol{g}^*$ and $\boldsymbol{g}$ are essentially identical.

Finally, we note that, since the α_i's are observed, one can always calculate $\frac{1}{CV^2(x_i)}$ for each study to find an upper bound on how much $\boldsymbol{g}$ and $\boldsymbol{g}^*$ are expected to differ. However, we note that in our analyses, details not shown, for practical sample sizes, $\boldsymbol{g}$ and $\boldsymbol{g}^*$will be very close.

Appendix D

We now consider more general disease models and study how a nonlinear effect of genotype count on disease risk impacts the efficacy of one-dimensional summaries of the posterior probability vector. Let x_i be the genotype of individual i and z_i be a vector of covariates. Let $\pi_i(x_i, z_i) = f(x_i, z_i)$ and $\pi_{0i} = f_0(z_i) = f(0, z_i)$. Note that the inclusion of covariates in the model suggests that π_{0i} may vary between individuals, unlike in previous appendices. Because x_i has only three states, any f is sufficiently described by a quadratic fit through the three points, and so without loss of generality we assume f is quadratic. Then:

$$\pi_i - \pi_{0i} = f(x_i, z_i) - f_0(z_i) = f'(0, z_i)x_i + f''(0, z_i)x_i^2.$$

There are two assumptions we can make to simplify the analysis: (i) f is sufficiently linear in x_i ($f''(0, z_i) = 0$) for all z_i values and (ii) that interactions between x_i and z_i are negligible ($f'(0, z_i) = f'(0)$ is free of z_i). A special form of (ii) occurs when we assume that the SNP and covariate effects are additively separable: $\pi_i = g(x_i) + h(z_i)$. Then $\pi_i - \pi_{0i} = g(x_i)$ and the problem reduces to the case without covariates. If we make both assumptions (i) and (ii), then the intuitions derived in *Appendices B* and *C* for a disease model where risk is linear in genotype count are expected to hold. The logistic model belongs to this class, thus explaining why dosage performs so well in this important case. To see this in the simple case of no covariates, let $\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \equiv f(x_i)$ and Taylor expand this expression about the dosage d_i:

$$\pi_i = f(d_i) + f'(d_i)(x_i - d_i) + f''(d_i)(x_i - d_i)^2.$$

One can show that $f''(d_i) = \beta_1^2 \pi_i (1 - \pi_i)(1 - 2\pi_i) \leq 0.1\beta_1^2$. A linear approximation to the logistic model results if we can justify ignoring the second order term. The upper bound already gives us some grounds for doing so. In addition, $f''(d_i)$ is near zero and thus negligible if $\pi_i \approx 0$

(low prevalence in a prospective cohort study), if $\pi_i \approx 0.5$ (approximately equal number of cases and control in a case-control study), or if the SNP effect is small.

We now move towards an analysis applicable towards a general disease model by first relaxing the assumption (ii), i.e., the effect of x_i does not depend on the value of z_i. Then $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \boldsymbol{V}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{B}\boldsymbol{x}$ where $\boldsymbol{B} = diag\big(f'(0, z_i)\big)$. The optimization problem for the non-centrality parameter becomes $\max_{\boldsymbol{g}\in\mathbb{R}^n} E(\boldsymbol{g}'\boldsymbol{B}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{B}\boldsymbol{g})$ subject to $\boldsymbol{g}'\boldsymbol{g} = 1$. If $\boldsymbol{B}$ is far from being proportional to the identity matrix, then this implies that the amount of "signal" we can expect from different individuals is different on average. Thus the scheme which treats all individuals equally is suboptimal.

We now relax assumption (i), linearity of the SNP effect. To optimize the noncentrality parameter on average, we solve we solve $\arg\max_{\boldsymbol{g}\in\mathbb{R}^n} E\big(\boldsymbol{g}'\boldsymbol{\mu}_{\boldsymbol{\beta}}\boldsymbol{\mu}_{\boldsymbol{\beta}}'\boldsymbol{g}\big)$ subject to $\boldsymbol{g}'\boldsymbol{g} = 1$. Recalling that $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ represents the vector of differences between the true disease probability and the null disease probability, it follows that from the Taylor series expansion that the $n \times 1$ vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ is given by $\big(\mu_{\beta i}\big) = (f'(0, z_i)x_i + f''(0, z_i)x_i^2)$. Following the results of *Appendix C*, the solution is the dominant eigenvector of $\boldsymbol{V}^{-\frac{\mathbf{1}}{\mathbf{2}}}\Big[E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)' + Cov\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)\Big]\boldsymbol{V}^{-\frac{\mathbf{1}}{\mathbf{2}}}$.

Using the same arguments from perturbation theory, $E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)$ is an essentially optimal one dimensional summary. Here, $E\big(\mu_{\beta i}\big) = E(f'(0, z_i)x_i + f''(0, z_i)x_i^2) = f'(0, z_i)d_i^{(1)} + f''(0, z_i)d_i^{(2)}$, where $d_i^{(j)} \triangleq E\left(x_i^{(j)}\right) = \alpha_{i1} + 2^j\alpha_{i2}$ gives what we call the j^{th} order dosage. We note that in most realistic situations $E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)'E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)$ will be large enough that the angle between the dosage vector, $E\big(\boldsymbol{\mu}_{\boldsymbol{\beta}}\big)$, and the optimal summary $\boldsymbol{g}$ will be essentially 0.

Note that $d_i^{(1)}$ is denoted d in the rest of this paper. The relative importance of the first-order versus the second-order dosage in our optimal summary is determined by the relative magnitudes of $f'(0, z_i)$ and $f''(0, z_i)$. If the effect on risk of the second allele differs significantly from that of the first, i.e. $f''(0, z_i)$ is large in magnitude, then the first-order dosage is an insufficient summary.

This is result is intuitive. The extent to which we need information beyond the dosage depends on the extent to which the disease model is non-linear. Note that $f'(0, z_i)$ and $f''(0, z_i)$ are not known, so implementing the optimal one-dimensional summary is infeasible for highly nonlinear risk models unless one is willing to make an educated guess about the relative degree of the second order effect. For example, if we believe that the second-order effect is some fraction ξ of the first-order effect, $f''(0, z_i) \propto \xi f'(0, z_i)$ for all z_i, then an optimal summary would be $d_i^{(1)} + \xi d_i^{(2)}$. Otherwise $\xi = \xi(z_i)$ may depend on the covariate value for each individual and the optimal summary for each individual would weight the first and second order dosages differently, $d_i^{(1)} + \xi(z_i) d_i^{(2)}$. Note that in most cases ξ is unknown and the degree of freedom lost by trying to estimate ξ may more than outweigh any efficiency gain from obtaining a better summary of x_i.

Appendix E

Suppose now that y_i is a quantitative trait. Let $Z = (\boldsymbol{z_1}, \dots, \boldsymbol{z_n})'$, where each z_i is a column vector of covariates for individual i of length l_z. Assume the normal linear model given by: $y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_i + \boldsymbol{\gamma' z_i} + \epsilon_i$ where γ is a constant length l_z column vector and $\epsilon_i \sim N(0, \sigma^2)$. Further let $X = (\boldsymbol{x}, Z)$, where $\boldsymbol{x}$ is $n \times 1$. The F-statistic testing $H_0 : \beta_1 = 0$ is

$$\frac{\boldsymbol{y}'(P_1 - P_0)\boldsymbol{y}}{\boldsymbol{y}'(I - P_1)\boldsymbol{y}/(n - l_z)}$$

Where $P_1 = X(X'X)^{-1}X'$ and $P_0 = Z(Z'Z)^{-1}Z'$. The noncentrality parameter of the statistic is an increasing function of:

$$\boldsymbol{\mu}'(P_1 - P_0)\boldsymbol{\mu} = \boldsymbol{x}'\beta_1[P_1 - P_0]\beta_1\boldsymbol{x}$$

$\boldsymbol{x}$ is unknown. Suppose we replace $\boldsymbol{x}$ with $\hat{\boldsymbol{x}}$ (accordingly P_1 with $\widehat{P_1}$). The non-centrality parameter for the F-statistic from using $\hat{\boldsymbol{x}}$ is $\boldsymbol{x}'\beta_1\left[\widehat{P_1} - P_0\right]\beta_1\boldsymbol{x}$**.** To maximize power, we again maximize this noncentrality parameter. That is, we seek to solve the following optimization problem:

$$\min_{\hat{x} \in \mathbb{R}^n} E\left[\boldsymbol{x}'\beta_1[P_1 - P_0]\beta_1\boldsymbol{x} - \boldsymbol{x}'\beta_1\left[\widehat{P_1} - P_0\right]\beta_1\boldsymbol{x}\right] = \min_{\hat{x} \in \mathbb{R}^n} E\left[\boldsymbol{x}'\beta_1\left[P_1 - \widehat{P_1}\right]\beta_1\boldsymbol{x}\right]$$

The first-order condition can be given as (in the expectation below, note that $\hat{\boldsymbol{x}}$ is fixed since it only depends on the covariates and hyperparameters governing the distribution of $\boldsymbol{x}$)

$$E\left[\beta_1^2\left(P_1 - \widehat{P_1}\right)\boldsymbol{x}\right] = E\left[\beta_1^2\left(I - \widehat{P_1}\right)\boldsymbol{x}\right] = \beta_1^2\left(I - \widehat{P_1}\right)E[\boldsymbol{x}] = 0$$

This says that if $\hat{\boldsymbol{x}}$ is optimal, then the column space of $\hat{\boldsymbol{x}}$ and Z must contain $E[\boldsymbol{x}]$. Thus one optimal solution is $\hat{\boldsymbol{x}} = E[\boldsymbol{x}]$, the dosage. This solution may not be unique just as the basis vectors of a vector space are not unique.

Figure 1 ROC Curve evaluating asymptotic significance of different approaches to summarizing posterior probability vectors

Caption: The asymptotic power of the dosage approach dominates the power of the test when using the mode, for all type I error rates. The figure illustrates the power of a test of case-control association for a SNP with MAF=0.10, odds ratio =1.28, and 1000 cases and 1000 controls. In this case the imputation r^2 was 0.60. The relatively low imputation r^2 explains why the dosage and mode are not performing better relative to the power of the test when using the true genotype.

Figure 2 Evaluating the power of the dosage and mode across different levels of imputation accuracy

Caption: Regardless of imputation accuracy, the dosage provides a more powerful choice of summary statistic than the mode. As expected, power increases as imputation accuracy increases. The figure illustrates the power of a test of case-control association for a SNP with MAF=0.10, odds-ratio=1.28, and 1000 cases and 1000 controls.

Figure 3 Evaluating the power of the dosage and mode across different values of the odds ratio

Caption For both small and large odds ratios, the dosage provided a more powerful alternative than the mode. The figure illustrates the power of a test of case-control association for a SNP with MAF=0.10, imputation r^2=0.6, and 1000 cases and 1000 controls.

Figure 4 Evaluating the power of the dosage and mode across different minor allele frequencies

Caption Across all minor allele frequencies, the dosage provided a more powerful alternative than the mode. The figure illustrates the power of a test of case-control association for a SNP with odds ratio of 1.25, imputation r^2=0.6, and 1000 cases and 1000 controls.