GLOGS: Appendices 1-6

Stephen Aaron Stanhope,[1] Mark Abney[1]

1 Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

Contact: sstanhop@bsd.uchicago.edu, (773) 489 1742

This appendix provides operational and implementation details as well as results omitted from the main paper for "GLOGS: A fast and powerful method for GWAS of complex diseases with risk covariates in related populations." It is broken into six sections.

Appendix 1 describes operational details of the GLOGS approach. These include initialization, null model estimation, and score statistical calculation. It can be used, in conjunction with the examples included in the software drop, to understand how to operate the software.

Appendix 2 provides technical details pertaining to model estimation. These include approximation of the likelihood function with a weighted sum over points in a cubature, and steps in the sample-information-resampling algorithm.

Appendix 3 provides derivations for the first and second derivatives of the likelihood function. These are used in both model estimation and score statistic calculations.

Appendix 4 provides explicit regression models corresponding to the various models used in our simulation studies, and Appendix 5 provides formal definitions of the various risk metrics we use to characterize our models.

Appendix 6 provides the results from our analysis of the sensitivity of GLOGS to choice of cubature size.

Appendix 1: Details of the GLOGS approach

We provide operational details of GLOGS in three steps: algorithm initialization; null model estimation; and score statistic calculation. For each step, we provide additional relevant commentary.

Step 1: Initialization

After covariate, phenotype and genotype data is collected and properly formatted, we perform a number of preliminary steps designed to support null model estimation.

1) We obtain initial parameter estimates $\hat{\beta}^0$ from a logistic regression that ignores polygenic effects. The purpose of $\hat{\beta}^0$ is to provide the numerical method for estimating mixed model parameters with an initial value for β.

2) We calculate a kinship matrix for the N individuals in the population either from a known pedigree (e.g. Abney 2009) or by estimating it from the observed genotypes as described in Thornton and McPeek (2010). Let the kinship matrix be denoted as Φ. We use Φ to calculate the cubature used for numerical integration, as described below.

3) We use the Sobol cubature generator provided by Joe and Kuo (2003, 2008) to obtain a cubature of C points distributed in the N-dimensional hypercube $[0,1]^N$. Let a point in the Sobol cubature be denoted as $\mathbf{u}_c$. We compute the multivariate normal cubature used for numerical integration by transforming individual points in the Sobol cubature according to $\mathbf{a}_c = \Sigma^{1/2} F^{-1}_{N[0,1]}(\mathbf{u}_c)$, where $\Sigma = 2\Phi$ and $F^{-1}_{N[0,1]}$ is a vectorized inverse probability transform to the standard normal. (E.g. the i^{th} element of $F^{-1}_{N[0,1]}(\mathbf{u}_c)$ is equal to $G^{-1}(\mathbf{u}^i_c)$, where G^{-1} is the inverse cumulative distribution function of a N[0,1] random variable.) Each point is assigned an initial weight equal to $w_c = C^{-1}$.

4) We select K polygenic variance values, $\sigma_k^0, k = 1, ..., K$, that are to be used in combination with β^0 as initial points for our numerical method.

We note that in principle, C multivariate normal random samples could be used in place of the Sobol sample generated in Step 3. We use the Sobol sample due to its theoretical advantages in comparison to a random sample, as well as its demonstrated performance on high dimensional integration problems similar to those we compute here. We further note that in practice, numerical optimization algorithms can show dependence on the initial parameter values. Our use of a logistic regression for obtaining initial covariate effects in combination with evaluating a number of possible initial values of polygenic variance is intended to minimize such dependency.

Step 2: Null model estimation

After the preliminary steps described in Step 1, maximum likelihood parameter estimates are computed based on each of the K initializing parameter sets $\{\hat{\beta}^0, \sigma_k^0\}$. Each of these calculations takes as inputs the covariate and phenotype data; $\{\mathbf{a}_c\}$ and $\{w_c\}$, $c = 1, ..., C$; $\{\hat{\beta}^0, \sigma_k^0\}$; and a set of algorithm control parameters $\{\delta, \epsilon\}$. We apply successive iterations of 1) a Newton-Raphson update (Press *et al* 1999, Chapter 9.6) on parameter estimates conditional on current cubature point weights followed by 2) cubature point reweighting and reassignment of point weights less than ϵ to zero, conditional on updated parameter estimates until convergence has been reached (see Appendix 2 for details).

Once complete, these calculations result in K combinations of maximum likelihood parameter estimates $\{\hat{\beta}^k, \hat{\sigma}^k\}, k = 1, ..., K$ and associated posterior cubature weights $\{w_c^k\}, c = 1, ..., C$. The final maximum likelihood parameter estimate and cubature weights are then taken to be those with the largest log likelihood amongst the K. Let them be denoted as

$\{\hat{\beta}^*, \hat{\sigma}^*\}$ and associated posterior cubature weights $\{w_c^*\}$ respectively.

Step 3: Score statistic calculation

After computing the maximum likelihood parameters and cubature weights, we conclude by using them to compute score statistics across all markers. Inputs into this step include the phenotype, covariate and genotype data; $\{\hat{\beta}^*, \hat{\sigma}^*\}$; and $\{w_c^*\}$. Once score statistics are computed, we perform genomic control to correct for any scaling errors and compute p-values from a χ^2 distribution with one degree of freedom.

Appendix 2: Logistic mixed model estimation

In describing our estimation procedure we concentrate on the general case (not estimation under the null model), with the understanding that inference under the null model can be accomplished by removing the genotype term. To approximate the likelihood integral in Eq. 2 as a weighted sum over points in a cubature, let C be the number of points in the cubature, $\mathbf{a}_c$ the c^{th} point, and w_c^i the weight of the c^{th} point at the i^{th} iteration. Let $\hat{\Theta}^i = \{\hat{\beta}^i, \hat{\gamma}^i, \hat{\sigma}^i\}$ be the i^{th} iteration parameter estimates. The approximated Eq. 2 evaluated at $\hat{\Theta}^i$ is then expressed as:

$$\ln(L^i(\hat{\Theta}^i)) \approx \ln\left(\sum_{c=1}^{C} \exp(l_{\mathbf{a}_c}(\hat{\Theta}^i)) w_c^i\right). \tag{1}$$

To solve for $\hat{\Theta}^{i+1}$ conditional on $\{w_1^i, ..., w_C^i\}$ we perform the following steps:

Step 1: Jacobian and Hessian computation

Let the Jacobian and Hessian of $\ln(L^i(\hat{\Theta}^i))$ be denoted as $D_\Theta \ln(L^i(\hat{\Theta}^i))$ and $D_\Theta^2 \ln(L^i(\hat{\Theta}^i))$ respectively. See Appendix 3 for component-wise derivative calculations.

Step 2: Perform Gauss-Newton parameter update

The Gauss-Newton updated parameter is:

$$\hat{\Theta}^{i+1} = \hat{\Theta}^i - D_\Theta^2 \ln(L^i(\hat{\Theta}^i))^{-1} D_\Theta \ln(L^i(\hat{\Theta}^i)) \tag{2}$$

Step 3: Check termination criterion

Termination is based on the Jacobian. If $||D_\Theta \ln(L^i(\hat{\Theta}^i))|| - ||D_\Theta \ln(L^i(\hat{\Theta}^{i+1}))|| < \delta$ then the estimation procedure has converged, and $\hat{\Theta}^i$ and $\{w_c^i\}, c = 1, ..., C$ are returned as the maximum likelihood parameter estimate and cubature weights.

Step 4: Update cubature weights

After solving for Θ^{i+1}, w_c^{i+1} are computed using Bayesian updating, followed by reassignment of weights less than ϵ to zero, and reweighting of the remaining points. Let

$$\begin{aligned} r_c^{i+1} &= \Pr(y|\mathbf{a}_c)\Pr(\mathbf{a}_c) && (3)\\ &= \exp(l_{\mathbf{a}_c}(\hat{\theta}^{i+1}))w_c^i. && (4) \end{aligned}$$

and $s_c^{i+1} = 0$ if $r_c^{i+1} < \epsilon$, and $s_c^{i+1} = r_c^{i+1}$ otherwise. Then, w_c^{i+1} is computed:

$$w_c^{i+1} = \frac{s_c^{i+1}}{\sum_{c=1}^{C} s_c^{i+1}} \tag{5}$$

Reassignment of cubature point weights near zero to be equal to zero is intended to speed computation times. Such points have only a minimal effect on likelihood scores and their derivatives, and if $w_c = 0$ then functions of $\mathbf{a}_c$ (e.g. $\exp(l_{\mathbf{a}_c}(\hat{\Theta}^i))$ in Eq. 5) do not need to be evaluated.

Appendix 3: Derivatives of the logistic mixed model

For both estimating parameters of the logistic mixed model as well as calculating score test statistics under the null model, derivatives of the approximated likelihood function in Eq. 5 must be evaluated. In this appendix we provide general forms for these derivatives. Ignoring iterations and use of estimated parameters, we want first and second derivatives of

$$\ln(L(\Theta)) \approx \ln\left(\sum_{c=1}^{C} \exp(l_{\mathbf{a}_c}(\Theta))w_c\right) \tag{6}$$

with respect to Θ. First derivatives take the form:

$$\frac{\partial}{\partial\Theta}\ln(L(\Theta)) = \frac{\sum_{c=1}^{C}\frac{\partial}{\partial\theta}l_{\mathbf{a}_c}(\Theta)\exp(l_{\mathbf{a}_c}(\Theta))w_c}{\sum_{c=1}^{C}\exp(l_{\mathbf{a}_c}(\Theta))w_c} \tag{7}$$

where

$$\frac{\partial}{\partial\Theta}l_{\mathbf{a}_c}(\Theta) = \sum_{n=1}^{N} z_n^{\Theta}(y_n - p_n), \tag{8}$$

$$p_n = \frac{\exp(\beta^T\mathbf{x}_n + \gamma g_{n,m} + \sigma\mathbf{a}_c(n))}{1+\exp(\beta^T\mathbf{x}_n + \gamma g_{n,m} + \sigma\mathbf{a}_c(n))}, \tag{9}$$

$\mathbf{a}_c(n)$ is the polygenic effect for the n^{th} individual, and z_n^{Θ} is the covariate for the n^{th} individual corresponding to the element of Θ under consideration (e.g. $z_n^{\Theta} = 1$ for a baseline effect term, and $z_n^{\Theta} = \mathbf{a}_c(n)$ for the polygenic effect). Second derivatives take the form:

$$\frac{\partial^2}{\partial\Theta\partial\Theta^T}\ln(L(\Theta)) = \frac{\sum_{c=1}^{C}(\frac{\partial^2}{\partial\Theta\partial\Theta^T}l_{\mathbf{a}_c}(\Theta) - \frac{\partial}{\partial\Theta}l_{\mathbf{a}_c}(\Theta)\frac{\partial}{\partial\Theta^T}l_{\mathbf{a}_c}(\Theta))\exp(l_{\mathbf{a}_c}(\Theta))w_c}{\sum_{c=1}^{C}\exp(l_{\mathbf{a}_c}(\Theta))w_c} - \frac{\partial}{\partial\Theta}\ln(L(\Theta))\frac{\partial}{\partial\Theta^T}\ln(L(\Theta)) \quad (10)$$

where

$$\frac{\partial^2}{\partial\Theta\partial\Theta^T}l_{\mathbf{a}_c}(\Theta) = \sum_{n=1}^{N} z_n^{\Theta} z_n^{\Theta^T}(p_n^2 - p_n). \quad (11)$$

Appendix 4: Simulation study models

Table 1 provides the logistic regression models corresponding to simulation models A-E in Table 2 of the main text. In our regression models, $g_{n,1}$ corresponds to the risky marker for the n^{th} individual, sex_n is a binary covariate corresponding to sex, and a_n is the sampled polygenic effect.

Model	$logit^{-1}(p_n)$
A	$-2 + 2g_{n,1} + 2a_n$
B	$-3 + 2sex_n + 2g_{n,1} + 2a_n$
C	$-4.5 + 3sex_n + 2g_{n,1} + 2a_n$
D	$-4.5 + 5sex_n + 2g_{n,1} + 2a_n$
E	$-4.5 + 8sex_n + 2g_{n,1} + 2a_n$

Table 1: **Simulation study models.**

Appendix 5: Risk calculations

An individual disease risk p is defined as:

$$p(x, g, a) = \frac{\exp(\beta x + \gamma g + \sigma a)}{1 + \exp(\beta x + \gamma g + \sigma a)}, \tag{12}$$

where x is a binary covariate, g is a biallelic marker, a is the polygenic effect, and β, γ, σ are effect sizes for fixed, genetic and polygenic covariates respectively. We are concerned with three quantities: 1) The relative risk associated with the binary covariate; 2) the relative risks associated with having one or two deleterious alleles; and 3) the sibling relative risk.

The relative risk associated with the binary covariate is equal to:

$$RR_x = \frac{\int p(1, g, a) f_g(g) f_a(a) dg da}{\int p(0, g, a) f_g(g) g_a(a) dg da} \tag{13}$$

where f_g and f_a are distributions for the number of deleterious alleles and polygenic effect, and we have integrated over both. In Table 1, we compute this as:

$$RR_x \approx \frac{\sum_{i=1}^{10000} p(1, g_i, a_i)}{\sum_{i=1}^{10000} p(0, g_i, a_i)} \tag{14}$$

where $g_i, i = 1, ..., 10000$ are sampled from a distribution taking the values $\{0, 1, 2\}$ with probabilities 0.5687, 0.3622 and 0.0691 respectively, and $a_i, i = 1, ..., 10000$ are sampled from a N(0,1) distribution. We note that the sampling distribution for g_i was determined using the average minor allele frequency, taken across all markers, of the Hutterite population that our simulation studies were based on.

Analogously, the relative risk associated with having 1 deleterious allele (and analogously for 2 alleles) is equal to and is computed:

$$\begin{aligned} RR_g(1) &= \frac{\int p(x,1,a)f(x,a)dxda}{\int p(x,0,a)f_x(x)f_a(a)dxda} && (15)\\ &\approx \frac{\sum_{i=1}^{10000} p(x_i,1,a_i)}{\sum_{i=1}^{10000} p(x_i,0,a_i)} && (16)\end{aligned}$$

where $a_i, i = 1, ..., 10000$ are again sampled from a N(0,1) distribution, and $x_i, i = 1, ..., 10000$ are sampled from a Bernoulli(0.485) distribution. This represents a 48.5% chance of an individual being male, which corresponds to the percentage of males in the Hutterite population under consideration.

The sibling relative risk is defined as the risk of two siblings being affected relative to that of two unrelated individuals. This is defined and is computed as:

$$\begin{aligned} SRR &= \frac{\int p(x^1,g^1,a^1)p(x^2,g^2,a^2)f_x(x^1)f_x(x^2)f_{gg}(g^1,g^2)f_{aa}(a^1,a^2)d\mathbf{x}d\mathbf{g}d\mathbf{a}}{\int p(x,g,a)f_x(x)f_g(g)f_a(a)dxdgda \int p(x,g,a)f_x(x)f_g(g)f_a(a)dxdgda} && (17)\\ &\approx \frac{\sum_{i=1}^{10000} p(x_i^1,g_i^1,a_i^1)p(x_i^2,g_i^2,a_i^2)}{\sum_{i=1}^{10000} p(x_i,g_i,a_i)\sum_{i=1}^{10000} p(x_i,g_i,a_i)} && (18)\end{aligned}$$

where in the numerator of our calculation, x_i^1 and $x_i^2, i = 1, ..., 10000$ are independently sampled from a Bernoulli(0.485) distribution, g_i^1 and $g_i^2, i = 1, ..., 10000$ take the values $\{0, 1, 2\}$ with probabilities 0.5687, 0.3622 and 0.0691 respectively, and a_i^1 and $a_i^2, i = 1, ..., 10000$ jointly from a multivariate normal distribution with standard marginals and covariance of 0.5. We note that the sampling procedure for g_i^1, g_i^2 is such that identity by descent is preserved for the sib pair while assigning one of the sibs marginal probabilities of having 0, 1 or 2 risk alleles as 0.5687, 0.3622 and 0.0691 respectively. In the denominator, x_i and a_i

are sampled as in the numerator, and g_i takes the values $\{0, 1, 2\}$ with probabilities 0.5687, 0.3622 and 0.0691 respectively.

Appendix 6: Effects of cubature size on p-values of associated markers.

To evaluate the effects of cubature size on the results of our analyses, we reanalyzed the datasets corresponding to model A in Table I using a Sobol cubature of 800000 points (2x the size of the initial study). After calculations based on the larger cubature size, $-\log_{10}$ transformed p-values for associated markers analyzed under both cubature sizes were plotted across simulations, and the power and type I error of the GLOGS procedure when using 800000 points was calculated and compared to that obtained when using 400000 points.

As reported in Table I, using a 400k-point cubature (first row) yielded an 88% detection power under a Bonferroni-controlled 5% test, with a marker-wise type I error rate of 2×10^{-5}. Using an 800k-point cubature (second row), the detection power and marker-wise type I error rates were 86% and 2×10^{-5} respectively, suggesting that increases in cubature size did not substantially change the statistical performance of our algorithm. To provide a visualization of the sensitivity of our results to cubature size, in Fig. A1 we compare $-\log_{10}$ p-values of associated markers from the 100 simulated datasets generated under model A across separate runs of the GLOGS procedure using 400k- and 800k-point cubatures. In Fig. A1, red lines represent 5% Bonferroni-controlled $-\log_{10}$ p-value thresholds, an individual point corresponds to the $-\log_{10}$ p-values of an associated marker from a particular dataset under 400k- and 800k-point cubatures (x- and y-axes respectively), and the identity line is colored in green. If the cubature size had a systematic effect on p-values of associated markers, we would anticipate a deviation of points from the identity line. However, the points are in fact scattered about the identity line, suggesting no systematic effect of cubature size on detection power.

Related to this test, we note the following regarding cubature choice and numerical integration. 1) For the analyses conducted in this paper, we use a transformed Sobol cubature. Besides the aforementioned theoretical advantages of doing so, the Sobol cubature has the practical advantage of not being a random sample. Rather, it is a quasi-random sample, and holding the dimensionality of the cubature constant, is determined solely by its size. This greatly simplifies sensitivity studies, such as conducted here. 2) As suggested, analysts can use random samples in the GLOGS procedure. We typically do not do so, for reasons previously discussed. Irrespective of whether or not the choice is made to use a Sobol quasi-random or random sample (cubature) as the basis for performing GWAS with GLOGS, we recommend utilizing as large as sample as possible given computational and time constraints. Doing so will give the most accurate results, as well as remove from consideration the issue of whether a larger cubature ought to have been used. 3) The size of the population under study is directly related to the size of cubature that can be feasibly used by the analyst, because each point in the cubature is a vector of length equal to the population size. This puts a practical limit, depending on available computational resources, on the population sizes that GLOGS can analyze. In our analyses of populations of approximately 800 individuals (more than double that of our simulation studies), a 400k-point cubature is nearly the largest usable without causing memory allocation errors. If we were to double that population size, we would anticipate only being able to use a 200k-point cubature.

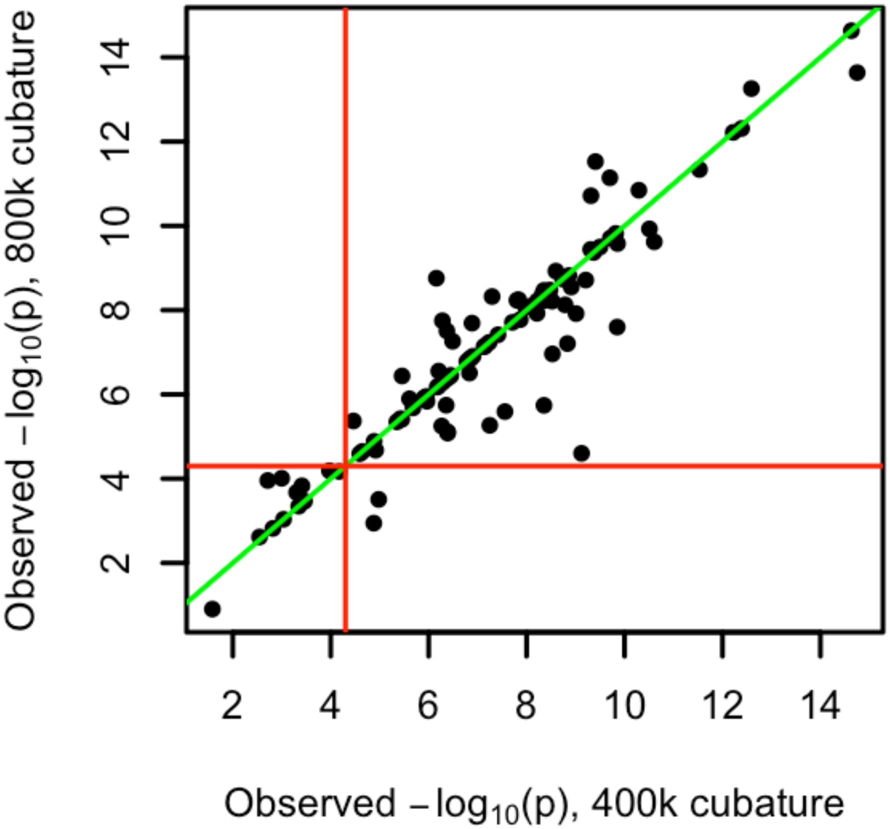

Figure 1: **Effects of cubature size on p-values of associated markers.** $-log_{10}$ p-values of associated markers from 100 simulations are compared across two analyses. The first uses a Sobol cubature of 400000 points, the second 800000 points. Red lines provide critical values for a Bonferroni-controlled 5% test.

References

Abney M. 2009. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. Bioinformatics 25: 1561-1663.

Joe S, Kuo F. 2003. Remark on Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator. ACM Transactions on Mathematical Software 29: 49-57.

Joe S, Kuo F. 2008. Constructing Sobol Sequences with Better Two-Dimensional Projections. SIAM Journal of Scientific Computing 30: 2635-2654.

Press W, Teukolsky S, Vetterling W, Flannery B. 1999. Numerical Methods in C. New York: Cambridge University Press.

Thornton T, McPeek M-S. 2010. ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. American Journal of Human Genetics 86: 172-184.