BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 11 2012, pages 1553–1554
doi:10.1093/bioinformatics/bts190

Genetics and population analysis

Advance Access publication April 19, 2012

GLOGS: a fast and powerful method for GWAS of binary traits with risk covariates in related populations

Stephen A. Stanhope* and Mark Abney

Department of Human Genetics, University of Chicago, 920 E. 58th St., Chicago, IL 60637, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Mixed model-based approaches to genome-wide association studies (GWAS) of binary traits in related individuals can account for non-genetic risk factors in an integrated manner. However, they are technically challenging. GLOGS (Genome-wide LOGistic mixed model/Score test) addresses such challenges with efficient statistical procedures and a parallel implementation. GLOGS has high power relative to alternative approaches as risk covariate effects increase, and can complete a GWAS in minutes.

Availability: Source code and documentation are provided at http://www.bioinformatics.org/~stanhope/GLOGS.

Contact: sstanhop@bsd.uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 5, 2011; revised on February 2, 2012; accepted on April 12, 2012

1 INTRODUCTION

Although many genome-wide association studies (GWAS) are of unrelated populations and do not include the effects of non-genetic predictors, it is also of interest to study populations with family structures or cryptic relatedness while incorporating the effects of covariates (Price *et al.*, 2010). Mixed model-based methods represent traits as a function of observed covariates and unknown random effects and are attractive for such problems because these effects are addressed in a direct and integrated manner. In such methods, the mean of the trait is affected by the observed covariates and the genetic marker being tested, whereas the random effects influence the trait's variance and correlation structure. Early implementations of the mixed model approach for quantitative traits include SOLAR (Almasy and Blangero, 1998). More recently, improvements in the efficiency of statistical methods, computing power and software engineering (Zhang *et al.*, 2009) have made the approach feasible for analyses of quantitative traits using genome-wide data, as systems such as TASSEL (Bradbury *et al.*, 2007) and EMMAX (Kang *et al.*, 2010) illustrate. Given these successes, further development of mixed model methods for GWAS is desirable, especially for binary traits.

In this article we describe Genome-wide LOGistic mixed model/Score test (GLOGS), a mixed model-based system for GWAS of binary traits in populations with related individuals. GLOGS performs a GWAS by (1) estimating parameters of a logistic risk model based on non-genetic covariates and a polygenic effect and (2) evaluating associations between markers and the disease using score tests. Simulation studies demonstrate that this approach yields comparatively high-statistical power, especially as covariate effects on risk increase. Additionally, GLOGS achieves a high degree of computational tractability through parallelization. GLOGS can complete GWAS in minutes, allowing investigations of a variety of possible covariates to be conducted in a timely manner.

*To whom correspondence should be addressed.

2 METHODS

2.1 Likelihood model and score test

For each individual $n=1,\ldots,N$ we observe a binary trait y_n, a vector of covariates x_n, and genotypes from M biallelic markers, where $g_{n,m}$ is the number of alleles of a particular type at the mth marker. We suppose that the probability of individual n carrying the trait (p_n) is influenced by an unobserved polygenic effect a_n, where $a_1,\ldots,a_N$ is $N(0,\Sigma)$ distributed, Σ accounting for relatedness and population structure.

Conditional on a_n we consider a logistic model relating covariates, the mth marker, and polygenic effects to risk; $p_n=\text{logit}^{-1}(\beta^T x_n+\gamma g_{n,m}+\sigma a_n)$ where β,γ and σ are the effects of the fixed, genetic and polygenic covariates. The related likelihood function is:

$$L(\beta,\gamma,\sigma)=\int_{R^N}\exp(l_a(\beta,\gamma,\sigma))h(a)da \quad (1)$$

$$l_a(\beta,\gamma,\sigma)=\sum_{n=1}^{N}y_n\ln(p_n)+(1-y_n)\ln(1-p_n) \quad (2)$$

where h is a multivariate normal $(0,\Sigma)$ distribution.

Based on Eqation (1), GLOGS performs GWAS by estimating parameters under the null model $H_0:\gamma=0$ via maximum likelihood, and calculating score statistics at each marker to evaluate H_0 versus $H_A:\gamma\neq 0$. If the mth marker has no effect, its corresponding score statistic is χ^2_1 distributed.

2.2 Implementation

Since Equation (1) and its derivatives do not have closed form solutions, we approximate them with weighted sums over a random sample or cubature. We calculate maximum-likelihood parameter estimates with a sampling–importance–resampling approach, and evaluate score statistics conditional on parameter estimates and posterior cubature weights. Details are provided in the Supplementary Material. GLOGS was implemented in C and MPI.

2.3 Simulation studies

We studied the performance of GLOGS with simulation studies of 369 individuals from the Hutterite population (Ober *et al.*, 1998, 2000), related through a 13-generation, 3028 member pedigree. We computed kinship coefficients Φ for these individuals with IdCoefs (Abney, 2009), supposed additive genetic effects and set $\Sigma=2\Phi$. We considered several risk models conditioned on a single marker, sex and the polygenic effect, representing a range of individual and relative risks (Table 1, explicit models are in the Supplementary Material). For each model, 100 sets of 1000 unlinked

Table 1. Simulation model properties and analysis results

Simulation model					WQLS		MQLS		GLOGS		
Model	Disease %	RR_{male}	RR_g	SRR	Type I error	Power	Type I error	Power	Type I error	Power	Run time (sec)
A	22.79	0	3.32, 7.35	1.88	0.006	83	0.006	90	0.002	88	107.98 + 0.15
B	23.81	2.78	3.02, 6.29	1.72	0.002	75	0.007	80	0.007	79	114.14 + 0.31
C	16.84	5.84	3.25, 7.34	1.86	0.007	63	0.008	74	0.005	78	119.74 + 0.19
D	30.39	12.69	2.14, 3.44	1.33	0.002	33	0.003	42	0.003	67	118.91 + 0.18
E	38.00	23.70	1.32, 1.60	1.05	0.004	2	0.002	5	0.005	27	121.35 + 0.17

RR_{male}, RR_g and SRR represent relative risks for males, 1 or 2 risk allele and sibling relative risks, respectively. Marker-wise type I errors and powers are expressed as percentages. Run times are given as time to perform model estimation plus perform 1000 marker score tests.

biallelic markers with 25% minor allele frequency were sampled by gene dropping. For each set of markers we sampled polygenic effects and trait statuses.

We evaluated the simulated data with GLOGS using a transformed 400 000 point Sobol cubature, computed run times, identified markers that rejected H_0 under a Bonferroni-controlled 5% test, and compared the results from GLOGS with those from WQLS (Bourgain *et al*., 2003) and MQLS (Thornton and McPeek, 2007). We additionally evaluated the sensitivity of GLOGS to cubature size by reanalyzing the data from model A using an 800 000 point cubature; results are provided in the Supplementary Material. These studies were performed using all cores of an Apple Mac Pro with two 3 GHz quad-core Intel Xeons and 8 GB of RAM, running OS X 10.5.8.

3 RESULTS

Table 1 shows marker-wise type I errors, powers and run times for our analyses. To obtain a 5% experiment-wise type I error rate, we used a Bonferroni controlled marker-wise type I error rate of 0.005%; no method has error rates significantly differing from this target. Detection power ranged from 2–83, 5–90 and 27–88% for WQLS, MQLS and GLOGS, respectively. For GLOGS, null model estimation was performed in under 2 min, and calculating score statistics for 1000 markers took under half-a-second.

4 CONCLUSION

In this article, we addressed mixed model-based GWAS of binary traits in populations of related individuals, where risk is affected by non-genetic factors. Our approach, GLOGS, calculates maximum likelihood parameter estimates for a logistic mixed model and then uses score tests to evaluate associations. We note that alternatives exist for both of these steps (Chen *et al*., 2011). Our use of maximum likelihood estimation is motivated by our use of the score test, which is based on maximum-likelihood estimation under the null model.

GLOGS improves upon other methods for GWAS of binary traits in related populations by offering direct support for risk covariates, in a system that is also powerful for studies that omit such covariates. This is demonstrated in Table 1, which suggests that GLOGS is roughly as powerful as the MQLS package when risk covariates have little effect, and more powerful for analyses with larger covariate effects. This is expected, as both methods are based on score tests and MQLS does not offer direct covariate support.

Additionally, GLOGS offers fast run times due to its use of the score test and parallelization of parameter estimation procedure. Typically, the total time required to perform a GWAS with GLOGS is primarily based on parameter estimation, which is done once per analysis with run time inversely proportional to the number of processors used. Score testing is comparatively fast. In Table 1, 1000 score tests were performed in ∼0.1% of the time of the parameter estimation step; increasing the number of markers 100-fold would yield an increased run time of 10%. In contrast, for MQLS and the methods discussed in Chen *et al*. (2011), a 100-fold increase in the number of markers would result in approximately a 100-fold increase in run time. Consistent with other methods, run times for GLOGS increase with the number of individuals studied.

We have used GLOGS in GWAS of over 800 highly related individuals and 250 000 markers, controlling for multiple binary, integer and continuous risk covariates. Such analyses take ∼20 min to perform with the computer hardware used in our simulation studies. A current version of GLOGS is provided at http://www.bioinformatics.org/∼stanhope/GLOGS.

ACKNOWLEDGEMENTS

We thank Mary Sara McPeek, Dan Nicolae, Carole Ober, Christopher King and the reviewers for their comments.

Funding: This study was funded by the National Institutes of Health grant R01 HG002899.

Conflict of Interest: none declared.

REFERENCES

Abney,M. (2009) A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, **25**, 1561–1663.

Almasy,L and Blangero,J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.

Bourgain,C *et al*. (2003) Novel case-control test in a founder population identifies P-selectin as an atopy susceptibility locus. *Am. J. Hum. Genet.*, **73**, 612–626.

Bradbury,P *et al*. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Chen,M.-H. *et al*. (2011) A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genet. Epidemiol.*, **35**, 650–657.

Kang,H. *et al*. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Ober,C. *et al*. (1998) Genome-wide search for asthma susceptibility loci in a founder population. *Human. Mol. Genet.*, **7**, 1393–1398.

Ober,C. *et al*. (2000) A second-generation genomewide screen for asthma-susceptibility alleles in a founder population. *Am. J. Hum. Genet.*, **67**, 1154–1162.

Price,A. *et al*. (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.

Thornton,T and McPeek,MS. (2007) Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.*, **81**, 321–337.

Zhang,Z. *et al*. (2009) Software engineering the mixed model for genome-wide association studies on large samples. *Brief. Bioinform.*, **10**, 664–675.