Bioinformatics Advance Access published July 8, 2010

METAL: Fast and efficient meta-analysis of genomewide association scans

Cristen J. Willer[1], Yun Li[1,2], Gonçalo R. Abecasis[1*]

[1]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, 48109

[2]Department of Genetics, Department of Biostatistics, University of North Carolina

Associate Editor: Prof. Burkhard Rost

ABSTRACT

Summary: METAL provides is a computationally efficient tool for meta-analysis of genome-wide association scans, which is a commonly used approach for improving power complex traits gene mapping studies. METAL provides a rich scripting interface and implements efficient memory management to allow analyses of very large datasets and to support a variety of input file formats.

Availability and Implementation: METAL, including source code, documentation, examples, and executables, is available at http://www.sph.umich.edu/csg/abecasis/metal/

Contact: Gonçalo Abecasis goncalo@umich.edu

1 INTRODUCTION

Meta-analysis is becoming an increasingly important tool in genome-wide association studies (GWAS) of complex genetic diseases and traits (de Bakker et al. 2008). Meta-analysis provides an efficient and practical strategy for detecting variants with modest effect sizes (Skol et al. 2007). We, and others, have used METAL for performing meta-analysis of GWAS to identify loci reproducibly associated with a variety of traits, such as type 2 diabetes (Scott et al. 2007; Zeggini et al. 2008), lipid levels (Willer et al, 2008; Kathiresan et al. 2009), BMI (Willer et al. 2009), blood pressure (Newton-Cheh et al. 2009) and fasting glucose levels (Prokopenko et al. 2009).

Meta-analysis of genome-wide association summary statistics, in contrast to direct analysis of pooled individual-level data, alleviates common concerns with privacy of study participants and avoids cumbersome integration of genotype and phenotypic data from different studies. Meta-analysis allows for custom analyses of individual studies to conveniently account for population substructure, the presence of related individuals, study-specific covariates, and many other ascertainment-related issues. It has been shown that meta-analysis of summary statistics is as efficient (in terms of statistical power) as pooling individual level data across studies, but much less cumbersome (Lin and Zeng 2009). Since genome-wide association studies routinely examine evidence for association at millions of directly genotyped and imputed SNPs across dozens or even hundreds of individual studies, it is important to use a fast and flexible tool to perform meta-analysis.

2 METHODS

The basic principle of meta-analysis is to combine the evidence for association from individual studies, using appropriate weights. METAL implements two approaches. The first approach converts the direction of effect and p-value observed in each study into a signed z-score such that very negative z-scores indicate a small p-value and an allele associated with lower disease risk or quantitative trait levels, whereas large positive z-scores indicate a small p-value and an allele associated with higher disease risk or quantitative trait levels. Z-scores for each allele are combined across samples in a weighted sum, with weights proportional to the square-root of the sample size for each study (Stouffer et al. 1949). In a study with unequal numbers of cases and controls, we recommend that the effective sample size be provided in the input file, where $N_{eff} = 4/(1/N_{cases}+1/N_{ctrls})$. This approach is very flexible and allows results to be combined even when effect size estimates are not available or the β coefficients and standard errors from individual studies are in different units. The second approach implemented in METAL weights the effect size estimates, or β coefficients, by their estimated standard errors. This second approach requires effect size estimates and their standard errors to be in consistent units across studies. Asymptotically, the two approaches are equivalent when the trait distribution is identical across samples (such that standard errors are a predictable function of sample size). Key formulae for both approaches are in Table 1.

Table 1. Formulae for meta-analysis

	Analytical Strategy	
	Sample Size Based	Inverse Variance Based
Inputs	N_i - sample size for study i P_i - p-value for study i Δ_i - direction of effect for study i	β_i - effect size estimate for study i se_i - standard error for study i
Intermediate Statistics	$Z_i = \varphi^{-1}(p_i/2) * sign(\Delta_i)$ $w_i = \sqrt{N_i}$	$w_i = 1/se_i^2$ $se = \sqrt{1/\sum_i w_i}$ $\beta = \sum_i \beta_i w_i / \sum_i w_i$
Overall Z-Score	$Z = \dfrac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$	$Z = \beta / se$
Overall P-value	$p=2\varphi(\lvert Z\rvert)$	

[*]To whom correspondence should be addressed.

3 RESULTS

3.1 Implementation

In implementing our software for meta-analysis, a primary consideration was to facilitate identification and resolution of common problems in meta-analysis. A secondary consideration was the ability to specify custom headers and delimiters so as to combine input files with varying formats generated from a variety of statistical packages. METAL tries to resolve or flag common problems that result from an inconsistent choice of allele labels or genomic strand across studies, or the presence of invalid p-values or test statistics at a subset of markers (due to numerical errors). METAL allows data to be filtered according to quality control measures, and can handle very large datasets (that typically total several GB in size) in workstations with a memory capacity not exceeding 2 GB.

3.2 Usage

METAL has been used extensively by many groups since its initial release in January 2008. This field testing enabled not only thorough debugging but improvements in error-detection methods. METAL can be run interactively or with a command script as input. Input files are processed one at a time and used to update intermediate statistics stored in memory. METAL implements Cochran's Q test for heterogeneity (Cochran, 1954) and the appropriate statistics can be calculated if requested by the user. METAL was designed for flexible formatting of input files, and allows users to customize labels for key columns, input field delimiters, and other characteristics of each input file. Information on genomic strand is used, if available, and - when it is unavailable - METAL automatically resolves strand mismatches for markers where strand is obvious (e.g, all SNPs except those with A/T and C/G alleles). METAL has an option to estimate a genomic control parameter (Devlin and Roeder 1999) for each input file and apply an appropriate genomic control correction to input statistics prior to performing meta-analysis. To facilitate the detection of allele labels that may have been misspecified by the user, which is critical for the correct determination of the direction of effect, METAL implements an option to output the mean, variance and minimum and maximum allele frequencies for each marker. METAL will track custom statistics, such as cumulative sample size, even when the standard error-weighted meta-analysis was performed. METAL can read gzipped files to allow for efficient use of disk space and optionally allows for subsets of markers to be analyzed. Full documentation of all options is available at http://www.sph.umich.edu/csg/abecasis/metal/.

3.3 Performance

METAL was written in C++ and is freely available for download. METAL compiles and runs on most Unix and Linux systems, and on Windows and Mac workstations. We recently performed a meta-analysis of GWAS for BMI (Willer et al. 2009). The analysis included 15 studies, each with association statistics at 2.2 – 2.5 million SNPs (average file size 225 MB), for a total of 36 million association statistics and a set of input files totaling 3.4 GB. This analysis required less than six minutes computing time and 790 MB of memory on a 2.83 GHz Intel processor. Run time scales linearly with the number of studies examined – a meta-analysis including 74 input files (each with > 2.5m SNPs) took 36 minutes and 1GB of memory.

ACKNOWLEDGEMENTS

The authors thank Michael Boehnke, Hyun Min Kang and Anne Jackson for reviewing early versions of this manuscript. We are also grateful to numerous collaborators in the GIANT Consortium, the Global Lipids Genetic Consortium, and the DIAGRAM Consortium for testing METAL and providing many useful suggestions.

Funding: GRA is supported by HG002651, MH084698, and HG005214. YL is partially supported by DK078150-03 and CA082659-11S1. CJW is supported by a Pathway to Independence Award K99HL094535

REFERENCES

Cochran, W.G. (1954). The combination of estimates from different experiments. Biometrics, 10, 101-129.

de Bakker, P. I., M. A. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri and B. F. Voight (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet, 17(R2), R122-8.

Devlin, B. and K. Roeder (1999). Genomic control for association studies. Biometrics, 55(4), 997-1004.

Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, G. M. Cooper, C. Roos, B. F. Voight, A. S. Havulinna, et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature Genetics, 40(2), 189-197.

Kathiresan, S., C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, et al. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet, 41(1), 56-65.

Lin, D. Y. and D. Zeng (2009). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol.

Newton-Cheh, C., T. Johnson, V. Gateva, M. D. Tobin, M. Bochud, L. Coin, S. S. Najjar, J. H. Zhao, S. C. Heath, S. Eyheramendy, et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet.

Prokopenko, I., C. Langenberg, J. C. Florez, R. Saxena, N. Soranzo, G. Thorleifsson, R. J. Loos, A. K. Manning, A. U. Jackson, Y. Aulchenko, et al. (2009). Variants in MTNR1B influence fasting glucose levels. Nat Genet, 41(1), 77-81.

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science, 316(5829), 1341-5.

Skol, A. D., L. J. Scott, G. R. Abecasis and M. Boehnke (2007). Optimal designs for two-stage genome-wide association studies. Genet Epidemiol, 31(7), 776-88.

Stouffer, S. A., Suchman, E. A , DeVinney, L.C., Star, S.A. , Williams, R.M. Jr (1949). Adjustment During Army Life. Princeton, NJ, Princeton University Press.

Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet, 40(2), 161-9.

Willer, C. J., E. K. Speliotes, R. J. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet, 41(1), 25-34.

Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet, 40(5), 638-45.