Mägi and Morris *BMC Bioinformatics* 2010, **11**:288
http://www.biomedcentral.com/1471-2105/11/288

BMC Bioinformatics

SOFTWARE Open Access

GWAMA: software for genome-wide association meta-analysis

Reedik Mägi*[1,2] and Andrew P Morris[1]

Abstract

Background: Despite the recent success of genome-wide association studies in identifying novel loci contributing effects to complex human traits, such as type 2 diabetes and obesity, much of the genetic component of variation in these phenotypes remains unexplained. One way to improving power to detect further novel loci is through meta-analysis of studies from the same population, increasing the sample size over any individual study. Although statistical software analysis packages incorporate routines for meta-analysis, they are ill equipped to meet the challenges of the scale and complexity of data generated in genome-wide association studies.

Results: We have developed flexible, open-source software for the meta-analysis of genome-wide association studies. The software incorporates a variety of error trapping facilities, and provides a range of meta-analysis summary statistics. The software is distributed with scripts that allow simple formatting of files containing the results of each association study and generate graphical summaries of genome-wide meta-analysis results.

Conclusions: The GWAMA (Genome-Wide Association Meta-Analysis) software has been developed to perform meta-analysis of summary statistics generated from genome-wide association studies of dichotomous phenotypes or quantitative traits. Software with source files, documentation and example data files are freely available online at http://www.well.ox.ac.uk/GWAMA.

Background

Genome-wide association (GWA) studies of hundreds of thousands of single nucleotide polymorphisms (SNPs), genotyped in samples of thousands of individuals, such as those undertaken by the Wellcome Trust Case Control Consortium [1], have proved successful in identifying novel common variants contributing moderate effects to a wide range of complex human traits (odds ratios greater than 1.2 for dichotomous traits or heritability of at least 1% for quantitative phenotypes). However, much of the genetic variation underpinning variation in these traits remains, as yet, unexplained. One natural way to increase power to detect rarer variants of more modest effect is to increase sample size. This can most readily be achieved through meta-analysis of multiple studies from the same or closely related populations, increasing the sample size to the order of tens of thousands. Such analyses have led to the identification of multiple, now established associations that would not otherwise have been identified in any individual study [2-4].

Meta-analysis of GWA studies has been greatly assisted by the development of imputation techniques [5,6] which predict genotypes not directly typed on available GWA genotyping products, but which are present on a dense reference panel of haplotypes, such as those available as part of the International HapMap Project [7] or the 1,000 Genomes Project [8]. With this approach, the results of GWA studies can be combined through meta-analysis of millions of SNPs, even if samples are interrogated with different GWA genotyping products.

The statistical methodology underlying meta-analysis is already well established [9], and freely available software packages provide routines for its implementation [10]. However, in the context of GWA studies, we face a number of unique challenges that make these existing tools impractical: (i) results are often combined for many studies for millions of SNPs, thus requiring memory efficient data manipulation; (ii) there may be over-dispersion of GWA test-statistics due to population structure, and between study variation, both of which must be

* Correspondence: reedik@well.ox.ac.uk
[1] Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
Full list of author information is available at the end of the article

accounted for in the meta-analysis; and (iii) computational difficulties in combining results obtained using different GWA genotyping products which may be aligned to different strands.

To address these challenges, we have developed the GWAMA (Genome-Wide Association Meta-Analysis) software to perform meta-analysis of summary statistics generated from GWA studies of dichotomous phenotypes or quantitative traits. The software incorporates tools to align studies to the same reference strand, irrespective of the GWA genotyping product, where possible, and optionally performs genomic control [11] of summary statistics to correct for population structure within each study, and potential variation between studies. The software also incorporates scripts for the generation of summaries of genome-wide meta-analyses including Manhattan and quantile-quantile (QQ) plots. Here, we demonstrate application of the GWAMA software to meta-analysis of 5 GWA studies, typed using different GWA genotyping products, but imputed at more than 2.3 million SNPs present on the International HapMap Project reference panels [7]. There are already several software packages available for meta-analysis and therefore comparison with some of the most widely used programs is part of current study.

Implementation

Consider a meta-analysis of *N* GWA studies, not necessarily typed using the same genotyping product or imputed to the same reference panel. We assume that studies have been filtered for appropriate quality control metrics to exclude poorly genotyped or imputed SNPs [12]. For each study, the following information is required for each good quality SNP: (i) the marker identifier; (ii) the allelic effect estimate and corresponding standard error (or an allelic odds ratio and 95% confidence interval in the case of a dichotomous trait); and (iii) the allele for which the effect has been estimated and the complimentary non-reference allele. Optionally, users may provide: (i) the frequency of the reference allele and the strand to which it has been aligned, which may aid alignment of AT/GC SNPs; (ii) the sample size contributing to the effect estimate; and (iii) an indicator to identify if the SNP has been directly genotyped in the study or imputed from a reference panel.

GWAMA begins by aligning all studies to the same reference allele at each SNP. If strand information is provided, a log file records potential misalignments and any corrections made based on the provision of reference alleles. Fixed effects meta-analysis is then performed for each SNP by combining allelic effects weighted by the inverse of their variance. The software performs tests of heterogeneity of effects across studies, and reports simple summaries of the direction of their effect in each to highlight potential inconsistencies in results. In the presence of heterogeneity of effects between studies, GWAMA can perform random-effects meta-analysis for each SNP by calculating the random-effects variance component. Graphical summaries of the results of the meta-analysis can be generated using the output of GWAMA, in conjunction with accompanying R scripts [10], provided that a map file containing SNP identifiers, chromosome and location are specified. A dense map file is provided with the GWAMA software which includes SNPs incorporated on a wide range of GWA genotyping products and variants present on the Phase 2 HapMap reference panel [7].

File formatting prior to meta-analysis

GWAMA is distributed with PERL scripts to format output from GWA association tools including PLINK [13] and SNPTEST [14]. The scripts extract the appropriate summary statistics from the output of these analysis packages, and allow subsequent filtering of results to exclude SNPs on the basis of minor allele frequency and/or number of called genotypes. However, we assume that studies have been otherwise filtered for appropriate quality control metrics to exclude poorly genotyped or imputed SNPs [12].

Study alignment and error trapping

GWAMA initially checks input data files for errors, such as negative values for odds ratios, and reports any issues to the log file. The study is then excluded from the meta-analysis for that SNP. The reference allele reported in the first study for each SNP is taken as reference, to which all allelic effects are then aligned (Table 1). If studies include estimates of the reference allele frequency, large discrepancies (more than 30%) are reported to the log file for manual checking. If strand information is not provided for studies, GWAMA assumes that alleles are aligned to the forward (+) strand of the NCBI dbSNP database. Strand misspecification is reported to the log file for all non- A/T or G/C SNPs, and are corrected before inclusion in the meta-analysis (Table 1). For A/T and G/C SNPs, strand errors cannot be detected, and all studies are assumed to have provided the correct alignment. However, to overcome potential strand issues for these SNPs, it is recommended that users provide reference allele frequency estimates, so that any large discrepancies between studies can be reported for manual checking.

Fixed-effects meta-analysis

Let β_{ij} denote the strand-aligned effect (log-odds ratio for a dichotomous phenotype) of the reference allele at the *j*th SNP in the *i*th study. The combined allelic effect across all studies at the *j*th SNP is then given by

Table 1: Example of alignment of allelic effects and error trapping for a single SNP in a meta-analysis of five studies of a dichotomous phenotype.

Study	Reported strand	Effect allele[1]	Other allele	RAF	Odds ratio (95% confidence interval)	Aligned allelic effect (standard error)	Comment
1	+	A	G	0.12	1.12 (1.07-1.16)	0.11 (0.02)	Allele A taken as reference effect allele.
2	+	G	A	0.85	0.92 (0.87-0.98)	0.08 (0.03)	Effect aligned to allele A.
3	-	T	C	0.12	1.06 (1.02-1.10)	0.06 (0.02)	Effect aligned to allele A on + strand.
4	+	T	C	0.13	1.07 (0.99-1.16)	0.07 (0.04)	Effect aligned to allele A on + strand. Strand error reported to log file.
5	+	A	G	0.87	0.95 (0.90-1.01)	-0.05 (0.03)	Large discrepancy in EAF reported to log file.

[1] Effects are aligned to the reference allele in the first study. Errors in the reported strand are recorded in the log file together with warnings regarding potential discrepancies in reported data between studies, for example the aligned reference allele frequency (RAF).

$$B_j = \frac{\sum_{i=1}^{N} \beta_{ij} w_{ij}}{\sum_{i=1}^{N} w_{ij}},$$

where $w_{ij} = [\mathrm{Var}(\beta_{ij})]^{-1}$ is the inverse of the variance of the estimated allelic effect in the *i*th study, obtained from the standard error (or 95% confidence interval of the odds ratio for a dichotomous phenotype). Note that if the *j*th SNP has not been directly genotyped or imputed as part of the *i*th study, $w_{ij} = 0$. The variance of the combined allelic effect across studies is given by $V_j = \left[\sum_{i=1}^{N} w_{ij} \right]^{-1}$. Furthermore, the statistic $X_j^2 = B_j^2 / V_j$ has an approximate χ^2 distribution with one degree of freedom, and this provides the basis of a test of association of the trait with the *j*th SNP over all studies.

Correcting for population structure

The presence of population structure in a GWA study, not taken account of in the analysis, can lead to over-dispersion of the corresponding association test statistics. One approach to combat this problem is to correct test statistics by the genomic control inflation factor. This factor is given by the median of the test statistics, divided by its expectation under the null hypothesis of no association, which is 0.456 in the context of an allelic-effect based analysis [11]. Users have the option to correct each study for potential population structure, hence the genomic control inflation factor is calculated separately for directly genotyped and imputed SNPs, denoted λ_{Di} and λ_{D^*i}, respectively, for the *i*th study [4,15]. The variance of each SNP in the study is then inflated by the relevant genomic control inflation factor so that $w_{ij} = \left[\lambda_{Ki} \mathrm{Var}\left(\beta_{ij} \right) \right]^{-1}$, where K is replaced by D or D^*, as appropriate. Furthermore, users have the option of correcting for between-study variation across the meta-analysis so that $X_j^2 = B_j^2 / \lambda V_j$. In this expression, λ is the genomic control inflation factor over all meta-analysed association test statistics, genome-wide.

Testing for heterogeneity between studies

To test for consistency of allelic effects across studies at the same SNP, GWAMA calculates two summary statistics of heterogeneity [16]. Cochran's statistic $Q_j = \sum_{i=1}^{N} w_{ij} \left(B_j - \beta_{ij} \right)^2$ provides a test of heterogeneity of allelic effects at the *j*th SNP, and has an approximate χ^2 distribution with N_j-1 degrees of freedom under the null hypothesis of consistency where N_j denotes the number of studies for which an allelic effect is reported. An alternative statistic, $I_j^2 = \left[Q_j - \left(N_j - 1 \right) \right] \big/ Q_j$, quantifies the extent of heterogeneity in allelic effects across studies, over and over that expected by chance, and is more robust than Q_j to variability in the number of studies included in the meta-analysis [17,18].

Random effects meta-analysis

In the presence of heterogeneity of allelic effects between studies, it is common to perform random-effects meta-analysis in order to correct the deflation in the variance of the fixed-effects estimate [19]. The random-effects variance component at the *j*th SNP is given by

$$\tau_j^2 = \max\left(0, \frac{Q_j - \left(N_j - 1 \right)}{\sum_i w_{ij} - \left(\sum_i w_{ij}^2 \big/ \sum_i w_{ij} \right)} \right),$$

and is used to inflate the variance of the estimated allelic effect in each study. The combined allelic effect across all studies at the SNP is then given by

$$B_j^* = \frac{\sum_{i=1}^{N} \beta_{ij} w_{ij}^*}{\sum_{i=1}^{N} w_{ij}^*},$$

where $w_{ij}^* = \left[\tau_j^2 + \mathrm{Var}\left(\beta_{ij} \right) \right]^{-1}$. The variance of the combined allelic effect across studies is given by $V_j^* = \left[\sum_i w_{ij}^* \right]^{-1}$. As in the fixed-effects meta-analysis, the statistic $X_j^2 = B_j^{*2} \big/ V_j^*$ has an approximate χ^2 distribution with one degree of freedom, and this provides the basis of a test of association of the trait with the *j*th SNP, allowing for heterogeneity of allelic effects between studies.

Output and analysis summaries

For each SNP, GWAMA will output a variety of summary information and statistics: (i) reference allele to which effects have been aligned and the corresponding non-reference allele; (ii) meta-analysis allelic effect estimate and standard error (or odds ratio and 95% confidence interval for a dichotomous phenotype); (iii) meta-analysis association test statistic, and corresponding *p*-value; (iv) heterogeneity test statistics *Q* (with *p*-value) and I^2; (v) heterogeneity summary, where each study is coded as '+' for increased effect of the reference allele, '-' for decreased effect of the reference allele, '0' for no effect of the reference allele, at a pre-specified significance threshold, and '?' if the study did not report an effect for the SNP. The output from GWAMA can be used with R scripts, supplied with the software, to generate QQ and Manhattan plots to summarise the genome-wide meta-analysis.

Results

To demonstrate the utility of GWAMA, we present the results of an example meta-analysis of 5 GWA studies of a simulated quantitative trait with directly typed and imputed genotypes at almost 2.4 million SNPs. Association summary statistics for each individual had previously been corrected for population structure, prior to meta-analysis. Statistical tests of association from the fixed-effects meta-analysis at each SNP were corrected for potential between-study variation on output using genomic control. The analysis was completed in just 3.5 minutes using a dedicated processor with 2.4 Gb memory. The data set used in this example is made available with GWAMA to test individual processor capabilities and potential limitations. To evaluate the memory capacity and program running time, we made additional testing with 20, 50, 100 and 200 genome wide datasets (each containing 2.4 million markers). The GWAMA program peaked with memory usage 4.8 GB, 8.2 GB, 14.6 GB, and 26.2 GB accordingly taking 10 min, 24 min, 53 min, and 1 hour 52 min to run.

Figure 1 presents QQ and Manhattan plots generated from GWAMA output using the summary R scripts released with the software. The QQ plot indicates that there is no evidence of population structure or between-study variation that has not been accounted in the analysis through genomic control. The Manhattan plot highlights two regions of association, on chromosomes 13 and 17, meeting genome-wide significance (SNPs in green have meta-analysis *p*-value less than 10^{-8}).

Discussion

There are currently several software packages designed for genome-wide meta-analysis of association test statistics including METAL [20], MetABEL [21] and META

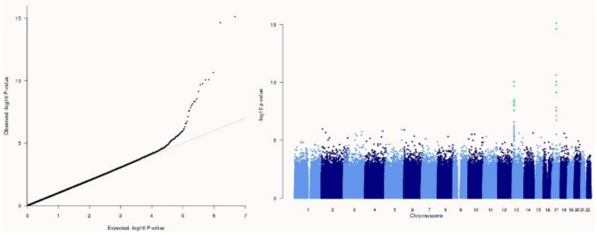

Figure 1 QQ and Manhattan plots generated from GWAMA output using the summary R scripts released with the software.

[22]. Table 2 presents a comparison of the key features of these software packages and GWAMA. The most important advantages of GWAMA over the existing packages are: (i) the distribution of supplementary scripts with the software to allow pre-processing of study summary statistic files generated by widely-used GWA analysis tools and production of graphical summaries to visualise the results of the meta-analysis; (ii) the calculation of two measures of heterogeneity of allelic effects between studies; (iii) the option to perform random-effects meta-analysis is the presence of heterogeneity; and (iv) genomic control correction of the association results of each study, and the meta-analysis overall, to allow for population structure.

Conclusions

In the coming months, we expect many more meta-analyses to be undertaken of increasing numbers of GWA studies of a wide range of phenotypes. With the imminent release of data from the 1000 Genomes project [8], we expect imputation to be performed at many millions of SNPs, generating ever larger sets of association summary statistics for analysis. GWAMA is designed to efficiently address the computational challenges of working with such large data-sets by filtering the necessary summary statistics from standard output files from GWA analysis software, as described above. Therefore, we expect that GWAMA will contribute to the identification of novel loci contributing effects to complex human traits in this exciting period of genetic research.

Availability and requirements

Project name: GWAMA

Project home page: http://www.well.ox.ac.uk/GWAMA

Operating system: UNIX (source code can be compiled with other platforms), Windows XP and newer

Programming language: C++, R, PERL

Other requirements: C++ compiler, optionally R version 2.9.0 or later with PNG support to generate graphics and PERL to run file formatting scripts

Licence: BSD

Any restrictions to use by non-academics: none

Authors' contributions

RM scripted and tested the software. APM provided statistical support and participated in error checking. Both authors wrote and approved the final manuscript.

Acknowledgements

The authors thank Dr. Cecilia Lindgren, Joshua C Randall, Dr. Eleftheria Zeggini and Dr. Inga Prokopenko for helpful suggestions regarding software implementation. Statistical analysis routines implemented in GWAMA utilised ALGLIB libraries made available through http://www.alglib.net.

Funding: Wellcome Trust (grant number WT081682/Z/06/Z); ENGAGE consortium, European Community Seventh Framework Programme (FP7/2007-2013); ENGAGE project, grant agreement HEALTH-F4_2007-201413; European Commission under the Marie Curie Intra-European Fellowship.

Table 2: Comparison of software packages for genome-wide meta-analysis of association summary statistics.

Software package	**METAL**	**MetABEL**	**META**	**GWAMA**
Pre-processing of GWA analysis files	No	*ABEL	SNPTEST	SNPTEST, PLINK
Strand flipping for aligning effect directions	Yes	Yes	Yes	Yes
Fixed effect analysis	Yes	Yes	Yes	Yes
Random effect analysis	No	No	Yes	Yes
Heterogeneity statistics (Cochran's *Q* statistic, I^2)	*Q*	No	*Q*, I^2	*Q*, I^2
Automated genomic control for population structure	Yes	Yes	Yes	Yes
Graphical visualisation of meta-analysis results	No	Forest plot	No	Separate scripts for Manhattan and QQ plots

Author Details

[1]Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK and [2]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Headington, Oxford OX3 7LJ, UK

Received: 17 November 2009 Accepted: 28 May 2010
Published: 28 May 2010

References

1. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**:661-678.
2. Zeggini E, Scott L, Saxena R, Voight B, Marchini J, Hu T, de Bakker P, Abecasis G, Almgren P, Andersen G: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.** *Nature genetics* 2008, **40(5)**:638-645.
3. Prokopenko I, Langenberg C, Florez J, Saxena R, Soranzo N, Thorleifsson G, Loos R, Manning A, Jackson A, Aulchenko Y: **Variants in MTNR1B influence fasting glucose levels.** *Nature genetics* 2008, **41(1)**:77-81.
4. Lindgren C, Heid I, Randall J, Lamina C, Steinthorsdottir V, Qi L, Speliotes E, Thorleifsson G, Willer C, Herrera B: **Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution.** *PLoS Genetics* 2009, **5(6)**:.
5. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes.** *Nature genetics* 2007, **39(7)**:906-913.
6. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81**:1084-1097.
7. The International HapMap Consortium: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851-861.
8. **The 1,000 Genomes Project** [http://www.1000genomes.org]
9. Lipsey MW, Wilson DB: **Practical meta-analysis.** Sage publications, California; 2001.
10. R Core Development Team: **R: a language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria; 2005.
11. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999:997-1004.
12. de Bakker P, Ferreira M, Jia X, Neale B, Raychaudhuri S, Voight B: **Practical aspects of imputation-driven meta-analysis of genome-wide association studies.** *Human molecular genetics* 2008, **17(R2)**:R122.
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, De Bakker P, Daly M: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *The American Journal of Human Genetics* 2007, **81(3)**:559-575.
14. **SNPTEST** [http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html]
15. Willer C, Speliotes E, Loos R, Li S, Lindgren C, Heid I, Berndt S, Elliott A, Jackson A, Lamina C: **Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.** *Nature genetics* 2008, **41(1)**:25-34.
16. Ioannidis J, Patsopoulos N, Evangelou E: **Heterogeneity in meta-analyses of genome-wide association investigations.** *PLoS One* 2007, **2(9)**:.
17. Huedo-Medina T, Sánchez-Meca J, Marín-Martínez F, Botella J: **Assessing heterogeneity in meta-analysis: Q statistic or I^2 index.** *Psychological Methods* 2006, **11(2)**:193-206.
18. Higgins JP, Thompson SG: **Quantifying heterogeneity in meta-analysis.** *Stat Med* 2002, **21**:1539-58.
19. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials* 1986, **7**:177-88.
20. **METAL** [http://www.sph.umich.edu/csg/abecasis/metal]
21. **MetABEL** [http://mga.bionet.nsc.ru/~yurii/ABEL/]
22. **META** [http://www.stats.ox.ac.uk/~jsliu/meta.html]

doi: 10.1186/1471-2105-11-288
Cite this article as: Mägi and Morris, GWAMA: software for genome-wide association meta-analysis *BMC Bioinformatics* 2010, **11**:288