Nucleic Acids Research Advance Access published January 12, 2012

Nucleic Acids Research, 2012, 1–8
doi:10.1093/nar/gkr1255

SURVEY AND SUMMARY

Comprehensive literature review and statistical considerations for GWAS meta-analysis

Ferdouse Begum[1], Debashis Ghosh[2], George C. Tseng[1,3,*] and Eleanor Feingold[1,3]

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, [2]Department of Statistics, Pennsylvania State University, University Park, PA 16802, and [3]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261, USA

Received August 23, 2011; Revised December 1, 2011; Accepted December 2, 2011

ABSTRACT

Over the last decade, genome-wide association studies (GWAS) have become the standard tool for gene discovery in human disease research. While debate continues about how to get the most out of these studies and on occasion about how much value these studies really provide, it is clear that many of the strongest results have come from large-scale mega-consortia and/or meta-analyses that combine data from up to dozens of studies and tens of thousands of subjects. While such analyses are becoming more and more common, statistical methods have lagged somewhat behind. There are good meta-analysis methods available, but even when they are carefully and optimally applied there remain some unresolved statistical issues. This article systematically reviews the GWAS meta-analysis literature, highlighting methodology and software options and reviewing methods that have been used in real studies. We illustrate differences among methods using a case study. We also discuss some of the unresolved issues and potential future directions.

INTRODUCTION

Genome-wide Association Studies (GWAS) test for statistical association between genotype and phenotype on hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) at a time in order to find genes that contribute to human diseases or non-disease traits. Early in the GWAS era, costs were high and sample sizes were small, but with technological advances prices have come down significantly and typical sample sizes are now in the thousands. Even with those large sample sizes, discoveries have been modest for many or most phenotypes studied because typical effect sizes are quite small, and many results do not appear to replicate in subsequent studies. As a result, most GWAS publications now involve multiple data sets in order to both reduce false positives and increase statistical power to find true positives. Often these multiple data sets are analyzed individually, or some of them are only used for 'in-silico replication' (i.e. only top markers from one data set are examined in the remaining data sets). There is growing recognition, however, that the most statistically robust and efficient analysis is a full-genome meta-analysis combining all studies and using all data at every marker. Meta-analysis provides optimum power to find effects that are homogeneous across cohorts, and at the same time can shed light on between-study heterogeneity (1–5). Going even further, many investigators are now forming mega-consortia of a dozen or more studies for increased statistical power. Meta-analysis thus has become a routine part of GWAS, and yet there remain unresolved issues about the most powerful and robust ways to use it. This article attempts to provide a comprehensive review of GWAS meta-analysis methods, practices and problems, with the goal of helping both applied and methodological researchers take the necessary next steps forward. In the next section we provide an overview of GWAS meta-analysis methods, and in 'Databases and software' we review databases and software. 'Literature review' summarizes the methods used in the literature, and 'Case study' presents our case study. Finally, in 'Complications and open questions' we discuss important open questions.

GWAS META-ANALYSIS DATA AND METHODS

It is fairly common for an individual investigator to perform GWAS on several different study populations and combine the results into a single report. If the genotyping is done for all studies together, data from the different populations can be directly combined

*To whom correspondence should be addressed. Tel: 412-624-5318; Fax: 412-624-2183; Email: ctseng@pitt.edu (termed 'mega-analysis'), and meta-analysis is not necessary. GWAS investigators generally turn to meta-analysis when scans are performed on different chips and/or when results from different investigators need to be combined and raw data cannot be exchanged for reasons of either confidentiality or proprietorship.

There has historically been some concern about the appropriateness of mega-analysis and even meta-analysis given the high level of heterogeneity among GWAS of the same trait. Sources of heterogeneity between studies can include different trait measurements and study designs, different ethnic groups, different environmental exposures, different genotyping chips, etc. For example, if two study populations have significantly different environmental backgrounds (say different diets in an obesity study), different genes may be relevant to the trait in the two populations (i.e. there may be gene × environment interaction). Another important source of heterogeneity is differing linkage disequilibrium patterns in different ethnic groups, so that even if the same variant is causal in both groups, the SNPs that are associated (in linkage disequilibrium) with it may differ from group to group. Recently, Lin *et al.* allayed some of these concerns. They showed both theoretically and by simulation that meta-analysis and mega-analysis have essentially equal statistical efficiency, and also that the efficiency of both approaches is fairly robust to between-study heterogeneity (6). Heterogeneity remains a concern, however, and we will discuss it further throughout the article (e.g. in the random effects model, case study and open questions).

Most GWAS meta-analysis uses relatively straightforward methods. *P*-values can be combined either with or without weights, or effect sizes can be combined in either fixed or random effects models. (See the companion paper on microarray meta-analysis for a more detailed exposition of the differences among these methods). Any of those methods can be applied either across all studies at once, or cumulatively as each study is added. Most GWAS meta-analysis takes a frequentist approach, but Bayesian hierarchical models can also be used, and are very well-suited to a cumulative approach (7). Table 1 lists the commonly-used GWAS meta-analysis methods and the source information that is required for each. The methods are described in a bit more detail below.

The simplest GWAS meta-analysis approach is to combine *P*-values using Fisher's method. The formula for the statistic is

$$X^2 = -2\sum_{i=1}^{k}\log(p_i)$$

where p_i is the *P*-value for the *i*th study. Under the null hypothesis, X^2 follows a chi-squared distribution with $2k$ degrees of freedom, where k is the number of studies. A major limitation of this method is that all studies are weighted equally, which is likely to be highly suboptimal when combining GWAS studies with different sample sizes. An additional problem is that the direction of effect of each SNP is not considered, so that studies with associations in opposite directions appear to strengthen each other rather than contradicting each other.

Table 1. Sources of information for different methods of meta-analysis

	Fisher's, *P*	Weighted *Z*	Fixed effect	Random effect
P-value	X	X		
Effect size			X	X
Direction of the effect size		X		
Sample size		X		
Heterogeneity estimate				X
SE of effect size			X	X

A major improvement over Fisher's method is a weighted *Z*-score method, in which *P*-values are transformed to *Z*-scores in a one-to-one transformation. The weighted *Z*-score method is more powerful and efficient than Fisher's method, and allows different weights for different studies (8). It also takes into account the direction of the effect at each SNP. The software METAL (9) implements the weighted *Z*-score method using the following formula:

$$Z = \frac{\Sigma_i z_i w_i}{\sqrt{\Sigma_i w_i^2}},$$

where the weight w_i = square root of sample size of the ith study, $Z_i = \Phi^{-1}(1 - \frac{pi}{2}) * (\mathit{effect\ direction\ for\ study\ i})$, and p_i is the *P*-value for the *i*th study. Note that the METAL paper has a typo in this formula but we have confirmed by testing the software that the formula shown above is in fact correctly implemented in the software.

The major alternative to combining *P*-values and/or *Z*-scores is to combine effect sizes (estimates). This can be done with either a fixed effects or a random effects model. Combining effect sizes is statistically more powerful than combining *Z*-scores, but it requires that the trait be measured on exactly the same scale in each study, with the same units, same transformations, etc. This may be achievable in a meta-analysis of a trait with highly standardized measurements, but there are many traits for which it is unlikely to be possible, for example alcohol or tobacco use. The difference between the fixed effects and random effects models is that fixed effects meta-analysis assumes that the genetic effects are the same across the different studies. Fixed effects models provide narrower confidence intervals and significantly lower *P*-values for the variants than random effects models (1,10–14). Both fixed effects and random effects models are briefly discussed below; details can be found in Nakaoka *et al.* (2009) (15).

For the fixed effects model, inverse-variance weighting is widely used, although other methods are also available. The weighted average of the effect sizes can be calculated as $\hat{\theta}_F = (\Sigma_i w_i\hat{\theta}_i)/(\Sigma_i w_i)$ and the variance of the weighted average of the effect size is $var(\hat{\theta}_F) = 1/\left(\sum{}_i w_i\right)$, where $\hat{\theta}_i$ is the logarithm of the *i*th case-control study effect, w_i is the reciprocal of the estimated variance of the effect size for the ith case-control study.

The random effects model assumes that the mean effect (of each SNP) in each study is different, with those means usually assumed to be chosen from a Gaussian distribution. The variance of that Gaussian distribution, and thus the amount of between-study heterogeneity, is estimated by the model. Thus the random effects model not only does not assume homogeneity of effect but is able to give an estimate of the degree of heterogeneity. The weight of each study incorporates the between-study variance of heterogeneity, which is expressed as τ^2, where

$$\tau^2 = (Q - (k - 1))/\left(\Sigma_i w_i - \left(\frac{\Sigma i w_i^2}{\Sigma i w_i}\right)\right).$$

The weight for the random effects model is calculated as $w_i^R = 1/(\frac{1}{w_i} + \hat{\tau}^2)$ and $Q = \Sigma_i w_i(\widehat{\theta_i} - \theta_F)^2$, Cochran's test statistic (16) follows a chi-squared distribution with $k-1$ degrees of freedom under the assumption of genetic homogeneity. Q is most widely used to check the between-study heterogeneity. But Q is underpowered when the number of studies is small. To overcome this problem, there are some other statistics available, such as H, R and I^2, defined as $H = \sqrt{Q/(k-1)}$, $R = \sqrt{\mathrm{var}(\hat{\theta}_\mathrm{R})/\mathrm{var}(\hat{\theta}_\mathrm{F})}$ and $I^2 = 100 * (Q - (K-1))/Q$, where $\hat{\theta}_R$ is the genetic effect under the random effects model. H, R and I^2 have some desirable characteristics such as being scale and size invariant (10,15). These statistics are calculated separately for each SNP, which leads to the interesting and unsolved question of whether or how one should make a genome-wide determination of heterogeneity.

In addition to these basic methods, almost any meta-analysis method in the statistical literature can be applied to GWAS, and some of the software packages discussed below do so.

DATABASES AND SOFTWARE

Most GWAS meta-analyses are assembled from consortia of investigators working on similar traits, but public databases are also used. The most important GWAS database is the NIH Database of Genotype and Phenotype (dbGaP), which is the repository for both raw data and results from most NIH-funded GWAS. There are also a number of databases that contain selected results from GWAS studies, some of which are suitable for inclusion in meta-analyses of targeted regions. GWAS Central is one of the oldest such databases, which started in 1998 under a different name. On 27 April 2011, it contained 708 studies. The Human Genome Epidemiology Network (HuGE Net) (http://www.hugenet.ca) also has a GWAS integrator webpage and contains a list of publications, hits, variants, disease and trait information etc. Like HuGE Net, The National Human Genome Research Institute (NHGRI) (http://www.genome.gov/gwastudies) maintains a catalog of published GWAS studies (17). Other available databases include the HKSC database with both bone mineral density (BMD) and fracture data (18) and the Millennium Genome Project (MGP) (https://gemdbj.nibio.go.jp/dgdb/), which has a repository of Japanese SNP(JSNP) data (19).

The statistical methods used for GWAS meta-analysis are very straightforward, and it is not difficult to implement them, but there are several software packages available that can make this easier and that integrate useful bioinformatics or visualization functions. The most widely used software is METAL (http://genome.sph.umich.edu/wiki/METAL_Program) (9). METAL implements two strategies, a weighted Z-score method based on sample size, P-value and direction of effect in each study, and an effect-size based method weighted by the study-specific standard error. The other most commonly used package is MetABEL, which is a component of the GenABEL suite in R. MetABEL implements a fixed effects model like METAL, and results can be shown with a visualization tool. A number of other packages are also in use, including META (http://www.stats.ox.ac.uk/~jsliu/meta.html). GWAMA (20) has useful auxiliary features that METAL, MetABEL, and META lack. PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/metaanal.shtml) (21) is a free, open-source software for GWAS analysis, which also has some meta-analysis tools to do fixed effects and random effects meta-analysis. MAGENTA (http://www.broadinstitute.org/mpg/magenta/) (22) can be used to test a specific hypothesis or to generate hypotheses, and it provides gene set enrichment analysis P-values and false discovery rate. Comprehensive Meta-analysis (CMA) (www.Meta-Analysis.com) Software (23) is a commercial package to do meta-analysis which works in a spreadsheet interface and also provides forest plots, which are useful for visualizing between-study heterogeneity (see case study). Review Manager (RevMan) (http://ims.cochrane.org/revman/about-revman-5) (24) is another package that does meta-analysis and provides results in tabular format and graphically. It also provides different kinds of reviews including intervention reviews, diagnostic test accuracy reviews, methodology reviews and overviews of the reviews. There are several STATA modules to perform meta-analysis, such as METAN (25), HETEROGI (25) and more specifically METAGEN (26) (http://bioinformatics.biol.uoa.gr/~pbagos/metagen) for genetic association studies. In *R*, a few other available packages for meta-analysis are Metafor (http://www.metafor-project.org/) (27), rmeta, and CATMAP. The Metafor package has different functions to calculate fixed, random and mixed effects along with moderator and meta-regression analysis and provides different kinds of graphical displays of results and data. Synthesis-view (https://chgr.mc.vanderbilt.edu/synthesisview) (28) is a visualization tool which can integrate multiple pieces of information across studies, such as P-values, effect sizes, allele frequencies etc. IGG3 (29) can integrate raw GWAS data from multiple chips and provide the input files for different imputation software, which can be used in meta-analysis later. Magi and Morris (2010) made a nice comparison of different features among a number of meta-analysis software packages (20).

One issue that is unique to GWAS meta-analysis is that SNPs may not be coded the same way in different data sets—the so-called ‘strand’ issue. Opposite coding of SNPs in different studies can cause what should be similar effects to look precisely opposite. This often occurs for only a small subset of SNPs (those with minor allele frequencies ~50%) and so can be very difficult to detect. Most of the meta-analysis software packages discussed above have varying bioinformatics features to resolve this problem, including METAL, MetABEL, META and GWAMA (20).

LITERATURE REVIEW

This review started with a search of GWAS meta-analysis using PubMed on 29 December 2010, which yielded 299 papers. After removing duplicates and irrelevant papers there were 249 GWAS meta-analysis papers (see complete searching and paper collection criteria in the companion paper). Figure 1 summarizes the number of papers by year of publication, illustrating the exponential increase between 2005 and 2010. Figure 2 summarizes the contents of the papers. One hundred and eighty-seven papers (75%) are full GWAS meta-analyses, while 58 papers (23%) are replication analyses on targeted loci (Figure 2A). Figure 2B shows that the majority of reports are biological applications (226 papers; 91%) while 10 papers (4%) are for novel methodology, 4 papers (1%) are databases and software, and 9 papers (4%) are review papers.

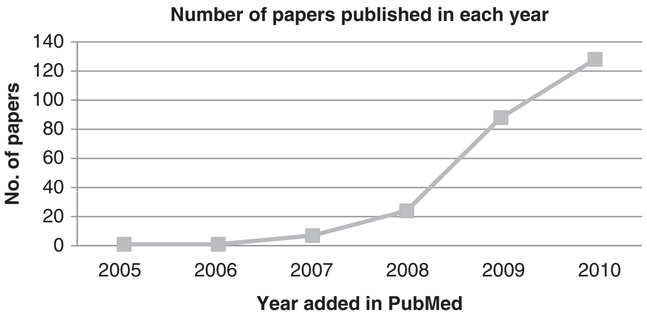

Figure 1. Number of GWAS studies by year of publication. Command used in PubMed search: [‘meta-analysis’(Title/Abstract)] AND [‘genome-wide association’(Title/Abstract)].

Figure 2C and 2D show the methods and software used. One hundred and sixty-four papers (80%) use fixed or random effects models, 28 (14%) combine weighted *Z*-scores from *P*-values, 6 (3%) use Fisher’s method, and 7 (3%) use direct data merging. For software packages, METAL (41 papers; 45%) and R packages (23 papers; 25%) are the most popular. Other software choices include PLINK (six papers; 7%); GWAMA (three papers; 3%); and GenABEL/MetABEL (five papers; 5%). Detailed information of the paper list and categorization to generate Figure 2 is available in the online Supplementary Data.

CASE STUDY

In this section, we present a simple case study that demonstrates some of the differences among GWAS

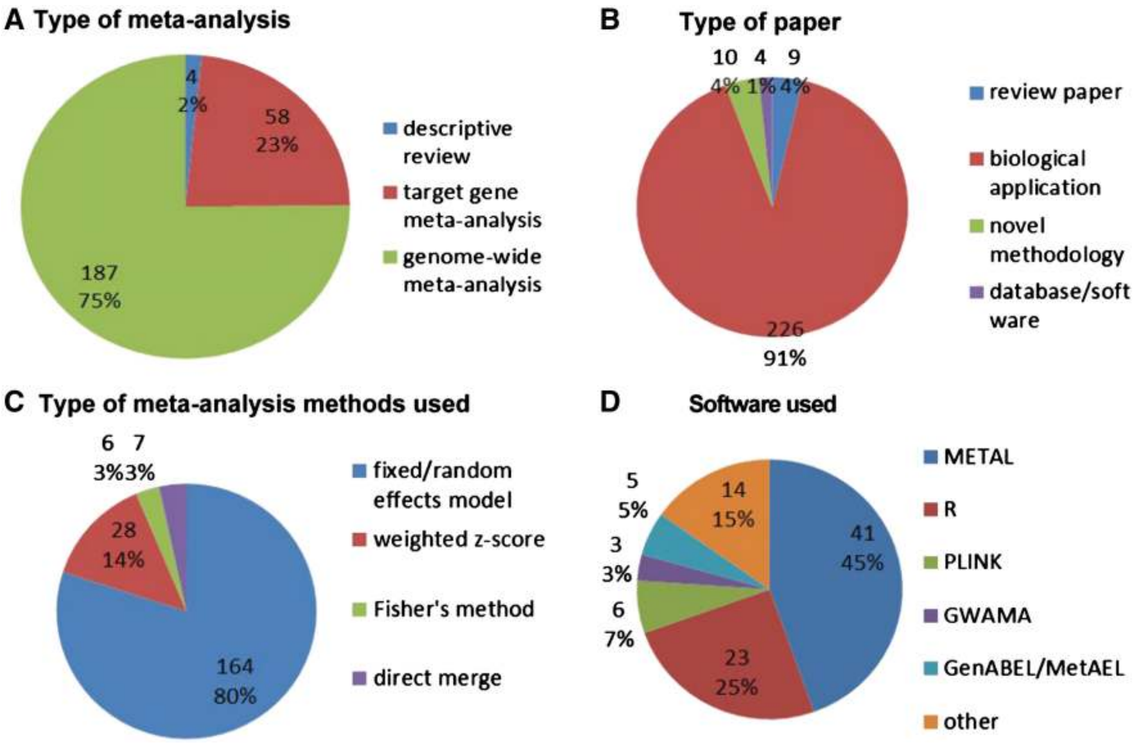

Figure 2. Summary of GWAS meta-analysis review: (**A**) type of meta-analysis; (**B**) type of paper; (**C**) type of meta-analysis method; (**D**) software used.

meta-analysis methods. Two data sets are included in this meta-analysis, which we label here as data set 1 and data set 2. The data sets are from different studies and different populations, but both were genotyped on the Illumina Human660-Quad Beadchip. The phenotype is total meiotic recombination across the genome, which has been of great interest in the genetics literature lately, with many new discoveries especially about the 'recombination hotspot gene' *PRDM9*. Meiotic recombination events for both parents in nuclear families were scored according to Chowdhury *et al*. (30) The gene *RNF212* is well-known to be associated with recombination (30–32), so we report results for four SNPs within this gene. Because the reported associations between *RNF212* and recombination differ in males and females, we consider males and females both separately and combined in our case study, which provides an illustration of how the different meta-analysis methods behave in the presence of heterogeneity. All the methods of meta-analysis for our case study were implemented by us in *R*.

Table 2 shows the results of our case study. The first four rows give the single-study *P*-values for each SNP in the four data sets (data set 1 male, data set 1 female, data set 2 male, data set 2 female). These are based on standard GWAS methods using linear regression for each SNP under an additive genetic model. No multiple comparisons correction was applied. The notable result is that all *P*-values are highly significant in the data set 1 males, but not in either set of females. In the data set 2 males, two of the SNPs have *P*-values of 0.01 and two are on the order of 0.20. Note that the sample size in data set 2 is much smaller than in data set 1, so even if the effects are the same in the two data sets we would expect larger *P*-values in data set 2.

Table 2. Case study results

	SNPs in RNF212			
	rs3796619	rs4974601	rs2045065	rs12645644
Study analysis				
Data set 1, *P*-value				
Male (n = 736)	1.4E − 6	1.4E − 6	1.7E − 6	1.8E − 6
Female (n = 736)	0.76	0.76	0.19	0.25
Data set 2, *P*-value				
Male (n = 174)	0.01	0.01	0.23	0.21
Female (n = 174)	0.15	0.14	0.82	0.82
Meta-analysis				
Fisher, *P*-value				
Male	2.7E − 7	2.7E − 7	6.2E − 6	5.9E − 6
Female	0.36	0.35	0.45	0.52
Combined	2.6E − 6	2.5E − 6	5.7E − 5	6.7E − 5
Weighted *Z*, *P*-value				
Male	2.35E − 8	2.35E − 8	6.87E − 7	6.34E − 7
Female	0.36	0.36	0.10	0.13
Combined	1.97E − 5	1.91E − 5	5.96E − 3	4.46E − 3
Fixed effect, *P*-value				
Male	1.7E − 8	1.7E − 8	7.0E − 7	6.3E − 7
Female	0.35	0.35	0.10	0.12
Combined	2.3E − 7	2.2E − 7	1.6E − 4	1.1E − 4
Random effect, *P*-value				
Male	1.7E − 8	1.7E − 8	1.7E − 1	1.5E − 1
Female	0.34	0.34	0.10	0.12
Combined	3.0E − 1	3.0E − 1	4.5E − 1	4.4E − 1

When the four meta-analysis methods are used to combine the two male data sets for the first two SNPs, they all perform reasonably well, but there are clear differences. Fisher's method has the lowest power (highest *P*-values), as would be expected because it is using equal weights for these two very different-sized sets. The highest power is found with both the fixed and random effects models; the similarity of these two methods for these two SNPs indicates that the fixed effects model fits well. For the third and fourth SNPs, the weighted *Z*-score method and the fixed effects model have better power than Fisher's method. The random effects model estimates a very large random component and gives a very high *P*-value for the SNP. This is probably an artifact caused by fitting a random effects model to just two data sets. Based on the biology, a fixed effects model is likely to be more or less correct for this phenotype, as long as only a single sex is included in the analysis.

In combining the female data sets, all four meta-analysis methods also behave similarly, reflecting the lack of significant association.

When all four data sets (males and females) are combined, we can clearly see the effect of the heterogeneity on the different meta-analysis methods. In general the fixed effects model retains good power to detect association despite our inclusion of some studies (the females) that have little or no effect, while the random effects model completely loses power because it is fitting an incorrect model of a Gaussian random effect. That is, our male and female effects are not the same, but they are not random either—what we actually have is a mixture of two fixed—effects models. We suggest that the typical situation in a GWAS meta-analysis is likely to be similar to this—a mixture of fixed effects rather than a true random effect—and thus that the random effects model may not be the most appropriate way to deal with heterogeneity in GWAS meta-analysis. This proposition clearly deserves further study, however.

One important way to visualize heterogeneity is with a forest plot, which shows the separate estimates and their confidence intervals for each study, and also shows the combination. Figure 3 is a forest plot for all four SNPs and all four populations in the case study; the overall effect shown in the forest plots is from the fixed-effects model. The *R* package 'rmeta' was used to generate the forest plots. These plots make it very easy to visualize some of the important features that the *P*-values only hint at, such as the fact that the two male populations are in fact quite consistent with each other despite the differing *P*-values, and the fact that the female effect is actually in the opposite direction (which is consistent with the recombination literature).

COMPLICATIONS AND OPEN QUESTIONS

GWAS meta-analysis is now widely used and in general has worked well to discover genetic effects that were not uncovered in individual studies. There are, however, some remaining barriers and open methodological issues.

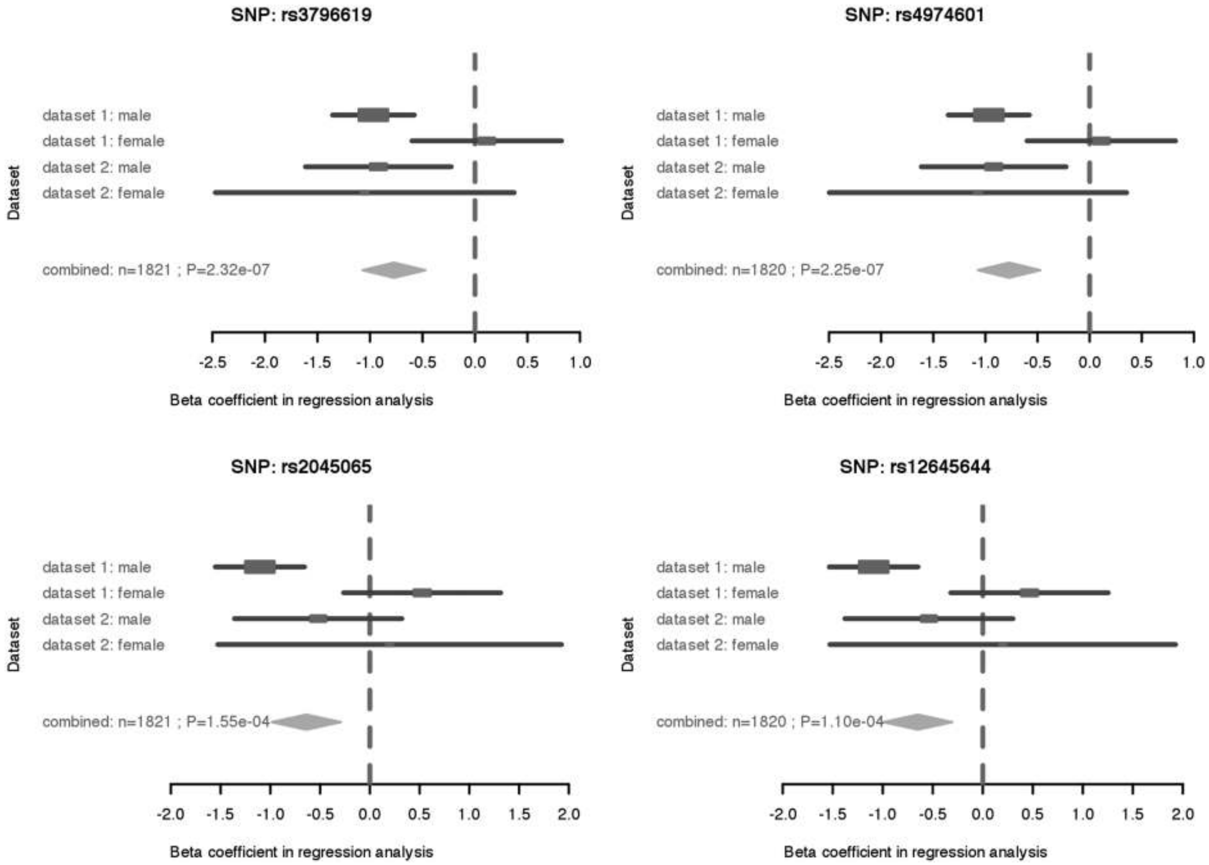

Figure 3. Forest plot of the selected SNPs.

Genotype data cleaning

Prior to meta-analysis, it is clearly important that all data sets undergo thorough standard GWAS data cleaning, such as filtering out 'bad' SNPs and samples using genotype call rates, tests of Hardy–Weinberg equilibrium (HWE), etc (33). What is not entirely clear is how important it is that the data cleaning steps and standards be the same across data sets. For example, can it cause problems if different genotype call rate cutoffs are used in different data sets? This has not been systematically studied to our knowledge. In genetic association studies for targeted SNPs, there have been three ways to deal with HWE: including all studies irrespective of the HWE tests (34), doing sensitivity analysis to verify differential genetic effects in subgroups (15,35–37), and excluding studies with statistically significant deviation from HWE(15,38). More recently, most large consortium meta-analyses have attempted to use consistent HWE cutoffs across studies, which is clearly the safest approach.

It is also not clear whether it is necessary or desirable to implement data cleaning steps that compare data sets to each other. The same SNP assay can behave differently on different chips, or even on the same chip in different batches, and thus it is common to scan data sets for SNPs with widely differing allele frequencies and eliminate them before combining. But if the data sets are from different ethnic groups, there will also be SNPs for which there are 'true' differences in allele frequency. It is not clear whether there is a way to distinguish the artifacts from the real differences, and thus it is difficult to recommend an ideal cleaning strategy. Similarly, HWE testing poses issues when data sets are combined (as discussed above), but it is probably clear that HWE tests on combined data sets would be unacceptably conservative. These issues are particularly important in the situation where different studies have different phenotype distributions (or, equivalently, different case : control ratios).

Imputation

When studies are genotyped on different chips, there may be very little overlap in the SNP sets, and thus direct SNP-by-SNP meta-analysis is impossible. For example, the overlap between the Illumina 550K SNP set and the Affymetrix 500K SNP set is only about 100K or 20% of SNPs. The standard solution to this problem is to impute the genotypes of all SNPs in all samples, and a variety of good methods is available for doing so (39). The problem this creates, which has not been carefully addressed in the literature, is that imputed genotypes have slightly higher error rates and variances than non-imputed genotypes. In general, if imputation is done carefully, the error rates are very low. Error rates can be higher, however, for areas of the genome with sparse SNP coverage or for ethnic groups that are not well represented in the data set that is used for imputation reference (usually HapMap or 1000 genomes). As with data cleaning above, this issue can be critical if different studies have different phenotype distributions. If two studies have different case:control ratios and one is genotyped and one imputed for a particular SNP, then there is a resulting difference between case and control variances, which can cause false positive results. Conversely, if one chip has very poor coverage of a region, then imputation will create 'genotypes' that actually convey very little information, in which case the meta-analysis can give false negative results because it is averaging in non-informative data sets. Some kind of regionally-smoothed meta-analysis may be the solution to this problem, but such methods have not been developed to our knowledge. In general, it is always advisable to check data quality of replicate results that are based predominantly on imputed data.

Choice of genetic models

In GWAS analysis, the basic association test can be based on an allele frequency comparison or on various statistical contrasts of genotype frequencies, for example an additive model, a dominant model, etc. The same model is used for each SNP, so usually something relatively robust such as the additive model is used (40). It is most desirable in meta-analysis to use the same model in each study, but in *post hoc* combinations of analyses that might not always be possible. To our knowledge, no one has studied the effect of such variation in association model on meta-analysis. Clearly it causes some level of effect heterogeneity that would, at least formally, violate a fixed effects model, though it would not fit a Gaussian random effects model either. Similar issues arise if different covariates or different methods for controlling for population stratification are used in different studies.

Between-study heterogeneity

As discussed above, between-study heterogeneity should probably be considered the norm in GWAS meta-analysis. Such heterogeneity is important to discover and report, since it can lead to important biological insights, for example differences in the genetic control of male and female recombination. The conventional wisdom in the statistical literature is that when heterogeneity is present or even likely, the random effects model is more appropriate than the fixed effects model. We suggest that this might not be the right approach for GWAS, because (i) the number of studies being combined is often not very large (leading to an imprecise heterogeneity estimate) and (ii) the form of the heterogeneity typically does not fit a Gaussian random effects model. We do suggest that forest plots are an important heuristic method for discovering and understanding heterogeneity, but we also propose that further work on random or mixed-effects models that are a better fit to GWAS data might improve analyses. For example, in our recombination example we know that males and females are likely to be different, so we could fit a model that explicitly has different fixed male and female effects.

CONCLUSION

As the GWAS literature moves away from artificial 'replication' and toward the more statistically optimal direct combination of all available data in a meta-analysis framework, it will be critical for investigators to understand the best methods for performing that meta-analysis. While good methods are already in use in most studies, there is room for improvement in many of the details discussed above. Many of the potential improvements are ideally addressed by planning studies in a coordinated manner from the beginning, but that is not always feasible. We still need improved methods for *post hoc* combinations of studies that may have significant heterogeneity in chip, study population, environmental exposures, association tests, etc. Looking even further ahead, all of the issues addressed above will need to be re-examined for meta-analyses of SNP data derived from sequencing studies, which will undoubtedly be appearing soon in journals throughout the field.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Drs Vivian Cheung and Mary Marazita for the use of their data in the case study. They also thank C Song, X Wang and G Liao for collecting and printing papers.

FUNDING

National Institutes of Health (NIH) (R01MH077159, RC2HL101715 to G.C.T.); NIH (R01HD38979, R01DE14899 to E.F. and F.B.); NIH (R01GM72007 to D.G.); Huck Institute for Life Sciences (to D.G). Funding for open access charge: University of Pittsburgh.

Conflict of interest statement. None declared.

REFERENCES

1. Higgins,J.P., Thompson,S.G., Deeks,J.J. and Altman,D.G. (2003) Measuring inconsistency in meta-analyses. *BMJ*, **327**, 557–560.
2. Ioannidis,J.P. (2007) Non-replication and inconsistency in the genome-wide association setting. *Hum. Heredity*, **64**, 203–213.
3. Thompson,J.R., Attia,J. and Minelli,C. (2011) The meta-analysis of genome-wide association studies. *Brief. Bioinformatics*, **12**, 259–269.
4. Thompson,S.G. (1994) Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*, **309**, 1351–1355.
5. Guerra,R. and Goldstein,D.R. (2010) *Meta-analysis and Combining Information in Genetics and Genomics*. CRC press, Taylor and Francis Group and a Chapman and Hall book, Florence, KY.
6. Lin,D.Y. and Zeng,D. (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.*, **34**, 60–66.
7. Zeggini,E. and Ioannidis,J.P. (2009) Meta-analysis in genome-wide association studies. *Pharmacogenomics*, **10**, 191–201.

8. Whitlock,M.C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.
9. Willer,C.J., Li,Y. and Abecasis,G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.
10. Higgins,J.P. and Thompson,S.G. (2002) Quantifying heterogeneity in a meta-analysis. *Stat. Med.*, **21**, 1539–1558.
11. Ioannidis,J.P.A., Patsopoulos,N.A. and Evangelou,E. (2007) Heterogeneity in meta-analysis of genome-wide association investigations. *PLOS One*, e841.
12. Lau,J., Ioannidis,J.P. and Schmid,C.H. (1997) Quantitive synthesis in systematic reviews. *Ann. Intern. Med.*, **126**, 820–826.
13. Lau,J., Ioannidis,J.P. and Schmid,C.H. (1998) Summing up evidence: one answer is not always enough. *Lancet*, **351**, 123–127.
14. Sutton,A.J., Abraham,K.R., Jones,D.R., Sheldon,T.A. and Song,F. (2000) *Methods for Meta-analysis in Medical Research*. John Wiley and Sons, Hoboken, New Jersey..
15. Nakaoka,H. and Inoue,I. (2009) Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse. *J. Hum. Genet.*, **54**, 615–623.
16. Cochran,W.G. (1954) The combination of estimates from different experiments. *Biometrics*, **10**, 101–129.
17. Kim,S.T., Cheng,Y., Hsu,F.C., Jin,T., Kader,A.K., Zheng,S.L., Isaacs,W.B., Xu,J. and Sun,J. (2010) Prostate cancer risk-associated variants reported from genome-wide association studies: meta-analysis and their contribution to genetic variation. *The Prostate*, **70**, 1729–1738.
18. Kung,A.W., Xiao,S.M., Cherny,S., Li,G.H., Gao,Y., Tso,G., Lau,K.S., Luk,K.D., Liu,J.M., Cui,B. *et al.* (2010) Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. *American J. Hum. Genet.*, **86**, 229–239.
19. Hirakawa,M.T.T., Hashimoto,Y., Kuroda,M., Takagi,T. and Nakamura,Y. (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.*, **30**, 158–162.
20. Magi,R. and Morris,A.P. (2010) GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, **11**, 288.
21. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American J. Hum. Genet.*, **81**, 559–575.
22. Segre,A.V., Groop,L., Mootha,V.K., Daly,M.J. and Altshuler,D. (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics*, **124**, 264–268.
23. Qu,H.Q., Bradfield,J.P., Li,Q., Kim,C., Frackelton,E., Grant,S.F., Hakonarson,H. and Polychronakos,C. (2010) In silico replication of the genome-wide association results of the Type 1 Diabetes Genetics Consortium. *Hum. Mol. Genet.*, **19**, 2534–2538.
24. Review Manager (RevMan) [Computer program]. (2011) *Version 5.1*. The Nordic Cochrane Centre, The Cochrane Collaboration, Copenhagen.
25. Patsopoulos,N.A. and Ioannidis,J.P. (2010) Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: a meta-analysis. *Ann. Rheumatic Diseases*, **69**, 561–566.
26. Wu,Y.W., Rong,T.Y., Li,H.H., Xiao,Q., Fei,Q.Z., Tan,E.K., Ding,J.Q. and Chen,S.D. (2010) Analysis of Lingo1 variant in sporadic and familial essential tremor among Asians. *Acta. Neurol. Scand*, **6**, e1001058.
27. Viechtbauer,W. (2010) Conducting meta-analyses in R with the metafor package. *J. Stat. Software*, **36**, 1–48.
28. Pendergrass,S.A., Dudek,S.M., Crawford,D.C. and Ritchie,M.D. (2010) Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Mining*, **3**, 10.
29. Li,M.X., Jiang,L., Kao,P.Y., Sham,P.C. and Song,Y.Q. (2009) IGG3: a tool to rapidly integrate large genotype datasets for whole-genome imputation and individual-level meta-analysis. *Bioinformatics*, **25**, 1449–1450.
30. Chowdhury,R., Bois,P.R., Feingold,E., Sherman,S.L. and Cheung,V.G. (2009) Genetic analysis of variation in human meiotic recombination. *PLoS Genet.*, **5**, e1000648.
31. Fledel-Alon,A., Leffler,E.M., Guan,Y., Stephens,M., Coop,G. and Przeworski,M. (2011) Variation in human recombination rates and its genetic determinants. *PLoS One*, **6**, e20321.
32. Kong,A., Thorleifsson,G., Stefansson,H., Masson,G., Helgason,A., Gudbjartsson,D.F., Jonsdottir,G.M., Gudjonsson,S.A., Sverrisson,S., Thorlacius,T. *et al.* (2008) Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science*, **319**, 1398–1401.
33. Laurie,C.C., Doheny,K.F., Mirel,D.B., Pugh,E.W., Bierut,L.J., Bhangale,T., Boehm,F., Caporaso,N.E., Cornelis,M.C., Edenberg,H.J. *et al.* (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.*, **34**, 591–602.
34. Minelli,C., Thompson,J.R., Abrams,K.R., Thakkinstian,A. and Attia,J. (2008) How should we use information about HWE in the meta-analyses of genetic association studies? *Int. J. Epidemiol.*, **37**, 136–146.
35. Salanti,G., Sanderson,S. and Higgins,J.P. (2005) Obstacles and opportunities in meta-analysis of genetic association studies. *Genet. Med. Official J. Am. College Med. Genet.*, **7**, 13–20.
36. Thakkinstian,A., McElduff,P., D'Este,C., Duffy,D. and Attia,J. (2005) A method for meta-analysis of molecular association studies. *Stat. Med.*, **24**, 1291–1306.
37. Zintzaras,E. and Lau,J. (2008) Trends in meta-analysis of genetic association studies. *J. Hum. Genet.*, **53**, 1–9.
38. Munafo,M.R. and Flint,J. (2004) Meta-analysis of genetic association studies. *Trends Genet.*, **20**, 439–444.
39. Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.
40. Kuo,C.L. and Feingold,E. (2010) What's the best statistic for a simple test of genetic association in a case-control study? *Genet. Epidemiol.*, **34**, 246–253.