The meta-analysis of genome-wide association studies

John R. Thompson, John Attia and Cosetta Minelli

Submitted: 21st September 2010; Received (in revised form): 23rd March 2011

Abstract

The pressure to publish novel genetic associations has meant that meta-analysis has been applied to genome-wide association studies without the time for a careful consideration of the methods that are used. This review distinguishes between the use of meta-analysis to validate previously reported genetic associations and its use for gene discovery, and advocates viewing gene discovery as an exploratory screen that requires independent replication instead of treating it as the application of hundreds of thousands of statistical tests. The review considers the use of fixed and random effects meta-analyses, the investigation of between-study heterogeneity, adjustment for confounding, assessing the combined evidence and genomic control, and comments on alternative approaches that have been used in the literature.

Keywords: *meta-analysis; genome-wide association studies; genome-wide significance; confounding; heterogeneity*

INTRODUCTION

Meta-analysis was originally developed as a method for pooling the results from a set of similar clinical trials. It was subsequently used to combine data from observational studies [1], then for combining genetic studies of candidate genes, and now meta-analysis is routinely used for pooling the results from genome-wide association studies (GWAS). Despite this broadening of the field of application, the methods used have changed very little, so more or less the same approach to meta-analysis is applied to genome-wide epidemiological data on hundreds of thousands of variants as were first developed for clinical trials of single outcomes.

Several textbooks are available that describe methods for analysing genetic epidemiology studies [2–4] and others describe general methods for meta-analysis [5–7]. Recently, several articles have reviewed the statistical techniques appropriate for the analysis of GWAS [8–11], but comparatively little theoretical work has been done on methods specifically intended for GWAS meta-analysis. Moonesinghe *et al.* [12] and Spencer *et al.* [13] consider issues of sample size and design, de Bakker *et al.* [14] consider the meta-analysis of imputed data and Zeggini and Ioannidis [15] discuss general meta-analysis issues in the context of GWAS. Recently, a special issue of the journal *Statistical Science* [volume 24(4) 2009] published a series of articles on the analysis of GWAS data including some relating to GWAS meta-analysis. In this review, we will look at the approaches that are currently being applied to meta-analyse genome-wide studies and comment on their merits. When assessing *ad hoc* adaptations of existing methods, it is important to remember that sensible procedures will often lead to the correct conclusion even if they are not theoretically ideal, especially when the genetic variant is clearly

Corresponding author. John Thompson, Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK. Tel: +44 (0)116 229 7269; Fax: +44 (0)116 229 7250; E-mail: trj@leicester.ac.uk

John Thompson is Professor of Genetic Epidemiology at the University of Leicester in England. He is a biostatistician with research interests that include the analysis and meta-analysis of genome-wide association studies, replication and Mendelian randomization. He has taught statistics and epidemiology at undergraduate and post-graduate levels.

John Attia is Professor of Medicine and Clinical Epidemiology at the University of Newcastle, Australia and the Hunter Medical Research Institute. His research focuses on the clinical applications of genetic association studies as well as meta-analysis of genetic data. He has taught clinical, molecular and genetic epidemiology at the undergraduate and graduate level.

Cosetta Minelli leads the Statistics Group in the Institute of Genetic Medicine at EURAC research (Bolzano, Italy). Her research focuses on the methodological issues of evidence synthesis in the field of genetics, in particular meta-analysis of candidate gene and genome-wide association studies.

associated with the disease. This may go some way to explaining why so many apparently contradictory approaches are used in the literature.

Throughout the review, the term genetic variant will be used, even though currently this almost always means single nucleotide polymorphism (SNP). Increasingly, however, meta-analyses will look at other variants, including haplotypes, copy number variants, genes and pathways. Although the detailed methods might change with the type of variant, the general principles that are considered here will remain much the same.

One of the features of genome-wide meta-analyses is that they can be undertaken for different reasons and each objective requires a slightly different form of analysis. When the motivation for the meta-analysis is not clear, it can be difficult to select an appropriate analysis and this difficulty is made worse when a single meta-analysis has several distinct components with different implicit aims. Our first task, therefore, is to consider the different forms of genome-wide meta-analysis and what they are trying to achieve. Following this, we consider some common, but controversial issues, namely; handling between-study heterogeneity and the use of fixed and random effects, adjustment for confounding, assessing the combined evidence and the use of genomic control.

TYPES OF GENOME-WIDE META-ANALYSIS

The most important distinction is between the use of a genome-wide meta-analysis for the discovery of new variants and its use for the replication of previous findings. While discovery analyses usually look across the whole genome, replication analyses concentrate on a limited number of pre-specified variants and as such have more in common with candidate gene studies. The most straightforward type of replication analysis occurs when a meta-analysis is used to assess the associations of previously suggested variants, but more contentiously, researchers sometimes run a GWAS followed by a replication study and then meta-analyse the replication data with the discovery data in order to capture what is sometimes called, the totality of the evidence.

The difference between discovery and replication parallels the difference between hypothesis generation and hypothesis testing and has consequences for the choice of analysis. Discovery is a process in which we screen the genome for good candidate regions. If a subset of genes are taken forward for further study, we will be concerned about the proportion of them that will eventually turn out to be null, the false discovery rate (FDR) [16]. We should be free to use whatever methods or information will improve the FDR, so often the discovery phase will take the form of an exploration of the data in which different types of analysis are tried. In contrast, in a replication, we seek to provide conclusive evidence for a candidate association, so it will be important to obtain a well-calibrated *P*-value or Bayes factor. To this end, the method of analysis and criterion for declaring replication should be specified in advance so as to avoid the possibility of the test being adapted in such a way that it exaggerates the significance.

Genome-wide meta-analysis for discovery

Within the field of discovery meta-analysis, two designs are commonly found in the literature. In the first, a consortium is formed of partners who have each conducted a GWAS of the same phenotype. The members of the consortium have the opportunity to work together to ensure the comparability of their quality control and primary analyses and to collaborate on more detailed follow-up analyses should interesting effects be observed [17]. The second design involves one or more primary GWAS that were initially intended to stand alone but which, perhaps for lack of power, find it difficult to obtain genome-wide significance and so go to the web in search of publicly available GWAS data that can be combined with their primary studies in order to obtain more precise results. Typically, such meta-analyses have fewer primary studies, have much less scope for secondary analyses and may even have difficulty ensuring the comparability of the quality control and methods of analysis. If the analogy with candidate gene meta-analysis can be relied on, then consortium-driven meta-analyses should be more reliable [18].

Genome-wide meta-analysis for replication

The replication of previously published hits using data from a genome-wide meta-analysis may lack the impact of new discoveries but it is still an important aim in its own right [19]. As well as confirming previous publications, it will give effect estimates that are free from the upward bias that results when a

study reports its own top hits, and what is more, the previous findings that are confirmed will give a degree of validity to the meta-analysis, since one might be doubtful of accepting discoveries from a study that could not replicate large known effects [20, 21]. Even a failure to replicate may be important, perhaps pointing to unsuspected interactions [22] or methodological problems with some of the studies. As more of the variants with large effects are identified and novel variants become harder to find, this type of confirmatory analysis is likely to become a larger part of any meta-analysis.

Replicating new discoveries

If researchers want to claim conclusive evidence for a new discovery, then it is important that an independent replication study is performed, because it may not be safe to rely on the discovery meta-analysis alone, no matter how strongly it may point to an effect [23, 24]. High statistical significance in a pre-specified test will effectively exclude the possibility that a false positive has been produced by sampling variation, but it cannot rule out the possibility that the finding is due to some bias in the design, analysis or conduct of the study.

Given the original discovery sample and a replication sample, possible strategies are to look for significant association in either:

- the replication sample alone;
- the replication sample combined with the discovery sample adjusted for multiple testing;
- the replication sample combined with the discovery sample unadjusted for multiple testing but judged against genome-wide significance.

In any of these options, a meta-analysis might be required to combine data from several replication samples and the second and third options are themselves a form of meta-analysis as they combine two data sources. The most popular strategies are the first and third, perhaps because the correct adjustment for the selection of top hits from the discovery phase in the second option is complex [25]. Skol showed some gain in power for option two over option one, but this gain is negligible unless at least 1% of SNPs are taken into replication or the replication sample is relatively very small. Arguably, only the first of these options is a true test of replication as the others are more accurately described as two-stage designs; that is to say, they are more efficient ways of performing a single study [26], and in a single study there remains the concern that the results may be influenced by bias. However, when the discovery study is itself a large meta-analysis including many studies, some of the need for an independent replication is removed as difference between the studies will often alert us to false positives due to biases.

Hybrid designs

Although, it is possible to identify broad classes of designs in published GWAS meta-analyses, there are actually many variations and there is no shortage of examples of studies that have combined two or more of the approaches. It is common practice, first to replicate findings from previous publications and then to explore the data to identify new candidate regions.

Genome-wide analyses involving non-association data

Finally, mention should be made of a type of discovery meta-analysis that has yet to have a large impact but which is likely to be used more in the future; that is to say, a meta-analysis that combines GWAS data with other types of information, perhaps biological characteristics of genes and SNPs throughout the genome from bioinformatical databases or from other experimental sources, such as expression studies. The aim here will be to improve the discovery phase by using the external information to weight the associations. For instance, associations on biologically relevant pathways identified using bioinformatical databases might be judged more worthy of follow-up than those in other regions [27]. Currently, this type of judgement is made informally but there is scope for incorporating the external data into the meta-analysis in a more structured and reproducible way, perhaps by creating prior weights for the variants [28] or by forming informative Bayesian priors. The extra data may not improve the measurement of the association, but it could make the selection of candidates for follow-up studies more reliable.

STAGES IN A META-ANALYSIS

Preliminary data processing

Meta-analysis starts with the quality control and analysis of the primary studies. Here, the plan must be to make all of the analyses as similar as possible to avoid causing unnecessary between-study heterogeneity.

Subjects should be excluded or the primary analyses adjusted if there is genetic or other evidence that they are drawn from a different population, or if there is a suggestion of cryptic relatedness [29]. Variants should be excluded if there is any suggestion that their genotype determination is unreliable; this decision might be based on high levels of missing genotype data, departure from Hardy–Weinberg equilibrium or allele frequencies in controls that do not correspond to those expected in that population. Usually, imputation will be used to extend the set of variants and to ensure that studies that have used different genotyping platforms are able to supply data on the same set of variants, in which case poor imputation quality will be grounds for excluding some variants [30].

The analysis of the individual primary studies needs to be kept as similar as possible, although sometimes differences are inevitable, for instance, when some samples have a family structure and others do not. The most popular primary analysis for a binary outcome is to calculate a per-allele odds ratio and its standard error using logistic regression or the Cochran–Armitage test. Per-allele analyses perform quite well even when effects are actually dominant or recessive [31], although robust tests may have better average properties [32].

Most collaborative meta-analyses ask the partners to deposit the results of the initial analyses with a co-ordinator who in turn makes them available to all partners. This enables the partners to retain control of their raw data, while placing few limitations on the form of the meta-analysis. Lin and Zheng [33] have shown that there is very little loss in statistical efficiency from pooling summary data when compared with a meta-analysis of individual patient data. The only drawback of working in this way comes when more detailed investigations are required, such as a haplotype analysis or an analysis of one SNP conditional on another, as the extra information will have to be requested from each partner. If legal and confidentiality issues can be overcome [34], depositing the raw data would speed up these secondary analyses.

Replication of previous discoveries

The large sample size of a meta-analysis may be needed to provide the power to successfully replicate small genetic effects. Replication should start with a review of the literature and of bioinformatical databases in order to identify candidate variants both for the particular trait under study and for associated traits, such as known intermediates or traits with a similar biology. In the replication of these candidates the form of the analysis needs to be specified in advance to avoid the possibility that researchers will make choices that alter the strength of the evidence. This pre-specification should not present a great problem because the information from the original reports will be available as a guide. Should some unexpected problem arise, such as unanticipated heterogeneity, then replication may be impossible to pursue and instead the unexpected findings should be described and possible causes of heterogeneity investigated in order to inform future studies.

Discovery: the initial screen

Discovery analyses are by their very nature exploratory and so it is not necessary to specify their final form in advance. Of course, as the analyses are not pre-specified they should not be taken as giving evidence of effect, but merely as providing suggestions of variants that can be tested in subsequent investigations. Typically, the discovery analysis will start with a simple, robust survey of the entire genome to identify variants that show either a large average effect, or a moderate effect with large between-study heterogeneity. Unadjusted fixed effects meta-analysis of per-allele effect estimates, which is based on the assumption that the effects are the same in all studies [5], combined with an assessment of between-study heterogeneity should suffice. Alternative screening procedures are to use a random effects meta-analysis, or to create a list from the overlap of top hits from each study, or to combine *P*-values rather than estimates. In a comparison of different approaches to discovery, the fixed effects analysis generally performed well, in the sense of placing true effects near the top of the combined list [35]. However, when heterogeneity is anticipated, a screen using a test that is more sensitive to individual studies with large effects may perform better [36].

Discovery: sensitivity analyses

Once an interesting variant has been identified, it is important to use sensitivity analyses to establish the robustness of the finding. Possibilities include investigating whether the effect extends over a chromosomal region or is confined to one variant, whether the results change if outlying phenotype measurements are excluded, whether the results depend critically on subjects whose genotype was difficult to call and whether the effect would be stronger under a

recessive or dominant genetic model. When large heterogeneity is found at a locus, it may be interesting to see whether all studies show an effect in that region, even if it peaks at a different specific variant. These analyses increase our confidence in the findings and may suggest lines for future study.

Discovery: secondary analyses

After establishing the robustness of the discovery, the next phase is the investigation of the mode of action of the variant. Often these analyses will require a re-analysis of the primary studies, for instance, to assess the importance of some variants adjusted for others, in order to see if the two sets act independently. Another secondary analysis of genetic is to adjust for measured intermediate phenotypes that lie on potential causal pathways, the idea being that if we adjust the analysis of each primary study for such a factor, then the effect of the variant on the outcome will be reduced. To distinguish such a reduction from random fluctuations in the estimates may well require the precision of a large meta-analysis. An extension of this approach that requires even larger sample sizes is the assessment of gene–gene or gene–environment interaction. Sub-group analyses based on primary studies that have measured that potential factor on each individual can look for differences in average effects between those sub-groups. Meta-regressions in which the effect size is regressed on some subject characteristic averaged over each study are usually easier to perform but have less precision than regressions based on individual level data and they may be biased [37, 38].

Bioinformatics is having an increasing impact on the types of secondary analyses that are possible. Meta-analyses now routinely look at patterns of linkage disequilibrium in interesting regions, many search for haplotypes across those regions and investigators are beginning to think in terms of sets of genes or pathways. Thus, if a good candidate lies in a gene on a particular pathway, it might be worth investigating variants in other genes on the same pathway even if their effect sizes are not as dramatically raised.

All of these secondary analyses require more detailed consideration of the data, which is why they are suited for use with a limited number of variants already identified in the initial screen or in previous studies. Never-the-less, these approaches have been attempted at the genome-wide level. For instance, it is possible to use data mining or machine learning approaches to search a genome-wide study for gene–gene interaction [39], or for interesting haplotypes [40]. While such genome-wide secondary analyses may occasionally be successful, it is likely that the successes will be swamped by false positives because of the number of possible combinations that need to be considered.

Replicating the new discoveries

As discussed earlier, replication requires confirmation of the new discoveries in an independent sample [23, 26]. Ideally, the replication sample should be drawn from a similar population using the same phenotype definition [24]. As the odds ratios of the top hits in the discovery sample will be upwardly biased [41], the actual power for replication can be a lot lower than it appears at first sight. Large replication samples will be needed and these should be planned for when the original meta-analysis is designed. A replication sample that is too small or which poorly matches the discovery sample leaves one unsure how to interpret variants those fail to replicate.

It is important to define in advance exactly what will constitute a replication, in particular whether a significant association must be demonstrated with exactly the same variant or whether another variant in the same region would be sufficient. The regional approach can have more power when linkage disequilibrium is high, although more variants will have to be genotyped and the test must correctly adjust for the extra comparisons [42]. Having established replication, it might be interesting to see if the effect generalizes to other, less similar, populations or to other related phenotypes.

FIXED EFFECTS, RANDOM EFFECTS AND THE IMPORTANCE OF HETEROGENEITY

Perhaps the most important choice facing the meta-analyst is whether to use a fixed or a random effects meta-analysis. In a fixed effects analysis, the underlying assumption is that at any location all of the studies have a common genetic effect and that the study-specific findings only differ from one another because of sampling variation, while in a random effects meta-analysis, it is supposed that each study population has its own size of genetic effect and that our aim is to estimate the average effect over all potential populations. This choice is not trivial because it can have a major impact on the

P-values and hence on the ranking of the variants. When the wrong model is used the *P*-values will be poorly calibrated in the sense that they will not achieve their nominal type-one errors. Thus a fixed effects meta-analysis applied in the presence of heterogeneity will tend to exaggerate the *P*-value, and a standard random effects meta-analysis [43] applied when the effects are actually common will be slightly conservative. The choice is made more difficult because it is quite possible that some variants will have the same effect in all populations, while other variants in the same scan have effects that vary across populations. A popular way out of this dilemma is to start with a fixed effects meta-analysis but to report the random effects meta-analysis when heterogeneity is found in a top hit. While sensible this option has its own problems; tests of heterogeneity are low powered so heterogeneity can be hard to establish when the number of studies is small, conditioning one test on another will distort the second test and cause its *P*-values to be mis-calibrated and it can lead to inconsistencies such as two variants in linkage disequilibrium, one analysed by fixed effects and the other by random effects. As Greenland [44] put it, 'if use of random effects makes a difference, the analysis is incomplete'.

The routine use of random effects meta-analysis would appear to be the safest option but it is not that popular, probably because of its potential conservatism and the importance of the *P*-value to publication; even in the absence of true heterogeneity some variants will appear to show heterogeneity by chance alone and the *P*-values for such variants will be unnecessarily penalized by a random effects analysis. Higgins *et al.* [45] give a thorough reappraisal of the strengths and weaknesses of random effects meta-analysis, although not in the context of a GWAS.

There is no perfect solution to the problem of choosing between a fixed and random effects analysis, but if a clear distinction is made between discovery and replication then the problem largely disappears. One of the aims of a discovery analysis is to find out whether there is any heterogeneity and having explored the discovery samples, the analyst will know what type of meta-analysis model to use in the replication. The *P*-value from the discovery phase becomes a screening tool, which may not have its nominal type-one error rate but this is of secondary importance, as in screening we are primarily interested in the sensitivity and FDR of whatever selection procedure we choose to employ [16, 46].

The obvious reason for performing a meta-analysis is to gain precision from having a larger total sample size. However, a second important reason is to describe and investigate heterogeneity [47–49]. This might either take the form of unanticipated between-study heterogeneity found in novel regions of the genome, or it might be that the meta-analysis was planned to look at heterogeneity in previously established variants.

Most genetic epidemiological studies measure average effects over a population and by doing so ignore the many individual variations that result from gene–gene and gene–environment interactions. So far these average effects have been the over-riding concern of GWAS meta-analyses, but variation in effect and the heterogeneity that results will inevitably become more important as our knowledge increases and more primary studies are conducted, especially those in people of non-European origin. Studies of subjects who are more thoroughly phenotyped will facilitate the meta-analysis of gene–environment interactions and perhaps lead to the discovery of the reasons behind some of the between-study heterogeneity.

When heterogeneity is found between studies, the first consideration must be to understand its possible causes and most importantly to distinguish between those causes that are the result of methodological differences and those that relate to true differences in the action of the variant. The process of measuring a genetic variant is complex and despite the best attempts to standardize the results, measurements made in different laboratories at different times will inevitably produce slightly different results. This effect can be magnified in a meta-analysis in which studies use entirely different genotyping platforms, impute using different software, or employ different quality control criteria. Meta-analyses that use downloaded data from the web are likely to be particularly affected by these problems.

Technical differences may produce a background noise of heterogeneity anywhere across the genome perhaps being worse for difficult to call variants such as those with low allele frequencies, while heterogeneity related to the action of the variant will cluster in regions where there is real genetic effect; after all, variants that are totally unconnected to the trait will be unconnected whatever the population under study. Heterogeneity that cannot be explained by methodological differences points to the presence of a gene–environment or gene–gene interaction

or possibly to the effect of variations in the phenotype definitions used in the studies. It is possible that the effects could move in different directions as in the flip-flop variants described by [50], but it is more likely that, if it were not for the impact of sampling variation, the effects would be in the same direction but with differing strength. If such effects are found then the next step would be to see if they extend over a region or are just restricted to a single variant. Isolated heterogeneity is much more likely to be due to technical differences, but a consistent pattern of heterogeneity over a region might indicate genes worthy of further study, even if their average effect does not reach conventional levels of genome-wide significance.

ADJUSTMENT FOR CONFOUNDING

In traditional epidemiological studies, it is common practice to adjust for age and gender and possibly for many other factors in order to remove their confounding effect on the relationship between the outcome and the variable under study. This practice of adjustment has crept into the analysis and meta-analysis of GWAS but often without being justified. Genetic variants are assigned to individuals before their birth so it is impossible for them to be changed by life-style or environmental factors that the individual experiences subsequently [51]. As a consequence, these factors cannot confound the association between the genetic variant and phenotype, even if they have a large impact of that phenotype. While one cannot totally rule out the possibility of confounding, it is generally unlikely and researchers that adjust for potential confounders should justify that choice.

Unnecessarily adjusting for a few non-confounders does little harm provided that they do not lie on the causal pathway between variant and phenotype and in some instances adjustment may even increase precision. However, if we adjust for a factor that turns out to lie on a causal pathway, or one that is highly correlated with a factor on a causal pathway, this could lead us to overlook an important variant. Combining unadjusted estimates would seem to be the natural default position for any discovery meta-analysis. Adjustment can then be seen as part of the secondary investigation of candidate variants. Meta-analyses that combine primary studies that have used different adjustments need to be viewed with caution as this can be a cause of heterogeneity.

A practical example of this problem is provided by McCarthy *et al*'s [10] review of the relationship between the fat-mass and obesity associated (FTO) gene and Type 2 diabetes. Overall the evidence suggests that FTO affects an individual's weight, which in turn increases their risk of diabetes. As a consequence of being on the same pathway, FTO showed an association with diabetes in population-based studies but failed to replicate in studies that controlled for weight by only recruiting lean subjects. This is not an example of confounding, because changing your weight will not alter your genotype and so adjustment for weight or BMI in a GWAS of diabetes cannot be justified on the grounds of confounding but would only be sensible if the researchers deliberately decided that they were not interested in variants that associated with diabetes via their effect on weight.

The one obvious exception to the principle that genetic associations are unlikely to be confounded is confounding by ethnicity, or population stratification as it has become known. Ethnicity might well affect the frequency of a variant and separately affect the phenotype. For this reason, population structure should always be investigated within each primary study and adjusted for when it is found [52–54]. For the same reason, adjustment for centre in multi-centre primary studies should also be routine.

ASSESSING THE COMBINED EVIDENCE

Genome-wide significance

One issue that causes much concern in the meta-analysis of GWAS is the definition of the level of evidence required to be confident that an effect is not due to chance. Like primary GWAS, meta-analyses usually define *P*-value thresholds at which they will declare a finding to be genome-wide significant, often values around the 5×10^{-7} or 10^{-8} are used [55–57], perhaps combined with a less stringent level of 10^{-5} or 10^{-6} at which variants will be considered interesting enough to warrant further investigation. Given the huge number of variants that are available for study, such low *P*-value thresholds are required to guard against the production of large numbers of false positives; however, they have the

drawback that such conservatism may cause many variants with small real effects to be overlooked.

The concept of genome-wide significance comes from viewing a GWAS, or the meta-analysis of GWAS, as a giant test of thousands or millions of independent hypotheses. However, due to the correlation between variants it is very difficult to specify separate hypotheses meaningfully and because of the many different ways in which a variant can act on a trait it is almost impossible to say in advance what statistical assumptions can be made when specifying the meta-analysis. As a consequence, the testing paradigm is unhelpful and it is better to view the discovery of new variants, not as a series of tests, but as an exploration, in which the *P*-values are a convenient screening tool rather than a measure of the strength of evidence.

Our ultimate aim is to judge whether a variant is null or not, a judgement that depends both on the *P*-value, which summarizes the evidence from the data under the null, and also on the plausibility of that null hypothesis. Unlike the test of an outcome in a clinical trial, our prior belief in a non-null effect of any single variant in a genome-wide study is very low, implying that we would need a much smaller *P*-value to convince us of a real effect. This has led some to argue that genome-wide studies should adjust *P*-values for the multiplicity of the whole genome rather than just for the variants being tested [57].

Of vital concern for a meta-analysis is whether *P*-values, as summaries of the evidence under the null, mean the same thing in all studies and in particular whether the selection of top hits or the exploratory use of different forms of analysis distorts their interpretation; what does a *P*-value mean if we only look at variants with *P*-values below some cut-off, or if we report a random effects *P*-value when we initially intended using a fixed effects meta-analysis? The answer to this question is critical, for if there is no common underlying scale a combined *P*-value of 2×10^{-9} based on meta-analysing a discovery *P*-value of 10^{-3} and a replication *P*-value of 10^{-7} cannot be assumed to mean the same thing as an identical combined *P*-value based on a discovery *P*-value of 10^{-7} and a replication *P*-value of 10^{-3}. Without a common scale, it is impossible to compare *P*-values from published reports that employ different selection criteria or analysis strategies and combining replication and unadjusted discovery samples would not be sensible.

Significance in the replication

Within the context of a pure replication the interpretation of a *P*-value is much more straightforward as there is no selection based on the same data set and the analysis will have been specified in advance. Most researchers use an adjustment for multiple testing but it could be argued that the key measure of evidence is the unadjusted *P*-value. When several variants are being tested for replication a secondary concern may be how many true effects will fail to replicate because of sampling variation, or whether some null variants will replicate by chance alone. It might be interesting to calculate the expected number of successful replications based on different assumptions about the true effects, but it is doubtful whether the use of Bonferroni adjustment, as it is commonly practised, will be very relevant in this context [58]. This adjustment seeks to control the chance of any of the replications appearing to be positive when they are really all null. Since it is very unlikely that all of the variants in a replication study will be null, the relevance of this calculation is doubtful. All it provides is the basis for choosing a rather conservative threshold [59].

Effect size estimates

It is well-known from empirical studies [41] and from theoretical considerations [60–63] that the odds ratio or any other measure of effect size of a variant selected, because it is a top hit, will be biased away from the null. This bias, which is sometimes called the winner's curse, arises because of sampling variation; imagine two genetic variants with the same true effect size and suppose that because of sampling the effect of one is over-estimated and the effect of the other is under-estimated, the SNP selected as a top hit will be the one that was over-estimated. Consequently the odds ratio of a top hit from a discovery study must be corrected for selection before it is combined with the odds ratio from a replication study [62], or used in a power calculation.

Bayesian alternatives and other extensions

Some GWAS have tried to avoid the limitations of *P*-values by adopting Bayesian methods. This approach is very attractive and would certainly also be useful in the context of a meta-analysis. The Wellcome Trust case-control consortium (WTCCC) calculated the Bayes factors of their main hits [64] and the coronary artery disease (CAD)

consortium adopted a similar approach but presented their results as posterior probabilities that a variant is null [65]. Both approaches gain in simplicity of interpretation at the expense of making the analysis conditional on some subjective prior assessments of the genetic effects [9, 66, 67]. It is interesting that both sets of researchers felt it necessary to include *P*-values as their main measures. Bayesian methods are also increasingly used for performing meta-analyses, although not as yet on GWAS data [68–70]. Other interesting developments concern attempts to define *P*-values across whole genes or even across whole genetic pathways [37, 71]. Such methods will undoubtedly become more important as biological knowledge accumulates in bioinformatical databases, but these *P*-values will suffer from the same difficulties of interpretation when used after selection.

GENOMIC CONTROL

It is now standard practice to look at the QQ-plots for each primary study and even to draw the QQ-plot of the final tests derived from the meta-analysis [72]. This type of plot compares an entire set of test statistics with those that would be expected if all variants were null. This plot was originally suggested as a way of detecting population stratification, however, deviations from the null pattern can arise for other reasons including, a noticeable proportion of positive variants, cryptic relatedness, genotyping error, or outliers in a continuous phenotype [53]. It is now common practice to divide all test statistics by λ, the ratio of the empirical to the null median test statistic [73].

As time has gone on, so the criteria for defining an acceptable λ have become stricter. At first a value of 1.10 in a primary study was acceptable, now anything over 1.05 is viewed as suspicious. Values of λ below one are often ignored although this is hard to justify. Great efforts are made to reduce λ by adjusting for ethnicity and relatedness, but studies that cannot reduce their λ to an acceptable level are likely to be down weighted in a meta-analysis by having their standard errors multiplied by root λ, or they may even be excluded entirely.

The λ is an average measure of inflation across the set of test statistics but it may be due to problems that affect some of the genetic variants and not others. Consequently, it should only be used after every attempt has been made to identify and eliminate the specific causes of the inflation, for instance by omitting phenotypic outliers, or adjusting for ethnicity. There seems little justification for penalizing the variants that pass these inspections because of problems with other variants that do not. Clayton *et al.* [53] extended this notion of treating variants separately by using a model that allowed variable down-weighting.

The problem of average adjustment for λ is even more critical in a meta-analysis because increasingly researchers are not only adjusting the *P*-values and standard errors from the primary studies, but also applying a second adjustment to the *P*-values from the meta-analysis. The logic here is not very clear because if the primary studies are acceptable then the inflation in the meta-analysis test statistics must be due to some aspect of the pooling; most likely heterogeneity between studies. It would seem more sensible to use a random effects analysis for the SNPs that show heterogeneity rather than to penalize all SNPs.

CONCLUSIONS

The growth in the use of GWAS has been so dramatic that there has been little time to consider how we should best use these exciting new data. The field has been dominated by the desire to be the first to publish evidence of a novel variant and the quality of the analysis has sometimes suffered as a result. Analysis methods seem to have been adopted not so much by detailed consideration as by imitation. For speed, researchers are sometimes content to use the same form of analysis as the last paper on that topic to appear in a prestigious journal.

The methods that are commonly used in the literature vary enormously from study to study and rarely have these methods been studied in sufficient depth to enable us to know their exact properties. None the less, the majority of the methods are very sensible and as such should identify the most important genes. It remains an open question whether these *ad hoc* methods make the best use of the data or whether they give an accurate picture of the strength of the evidence.

We take the view that a meta-analysis of GWAS should be viewed partly as an opportunity to replicate candidate variants using formal statistical tests and partly as an opportunity to discover new variants or to explore the way that variants operate. The discovery phase is essentially a simple screen of the

whole genome followed by a number of sensitivity and secondary analyses informed by bioinformatics. Conclusive evidence of a genetic association comes when the discoveries are confirmed in an independent replication study.

Key Points

- Pressure to publish novel genetic associations has meant that meta-analysis has been applied to genome-wide association studies without the time for a careful consideration of the methods that are used.
- Discovery GWAS meta-analyses should be treated differently from meta-analyses intended to validate a previous discovery. In particular, it is better to think of the discovery phase as an exploratory screen rather than as a set of statistical tests.
- In the discovery phase, the meta-analysis should include investigations of heterogeneity and the robustness of the findings, as well as secondary analyses aimed at suggesting likely causal mechanisms. All analyses should be informed by bioinformatical data.
- Adjustment for confounding and the use of genomic control have become widely accepted as standard practice in GWAS meta-analysis but their use is not always justified and deserves more careful consideration.

References

1. Egger M, Schneider M, Smith GD. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998;**316**: 140–14.
2. Elston RC, Olsen JM, Palmer L. *Biostatistical Genetics and Genetic Epidemiology*. Wiley, Chichester, 2002.
3. Thomas DC. *Statistical Methods in Genetic Epidemiology*. Oxford University Press, New York, 2004.
4. Ziegler A, König IR. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley, Weinheim, 2010.
5. Sutton AJ, Abrams KR, Jones DR, *et al*. *Methods for Meta-analysis in Medical Research*. Wiley, Chichester, 2000.
6. Borenstein M, Hedges LV, Higgins JPT, *et al*. *Introduction to Meta-analysis*. Wiley, Chichester, 2009.
7. Sterne J. *Meta-analysis in Stata*. Stata Press, College Station, Texas, 2009.
8. Wang WYS, Barratt BJ, Clayton DG, *et al*. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6**:109–18.
9. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;**7**:781–91.
10. McCarthy MI, Abecasis GR, Cardon LR, *et al*. Genome-wide association studies for complex traits: consensus uncertainty and challenges. *Nat Rev Genet* 2008;**9**: 356–69.
11. Ziegler A, König IR, Thompson JR. Biostatistical aspects of genome-wide association studies. *Biom J* 2008;**50**:8–28.
12. Moonesinghe R, Khoury MJ, Liu T, *et al*. Required sample size and non-replicability thresholds for heterogeneous genetic associations. *Proc Natl Acad Sci USA* 2008;**105**:617–22.
13. Spencer CCA, Su Z, Donnelly P, *et al*. Designing genome-wide association studies: Sample size power imputation and the choice of genotyping chip. *PLoS Genetics* 2009;**5**:e1000477.
14. de Bakker PIW, Ferreira MAR, Jia X, *et al*. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008;**17**:R122–8.
15. Zeggini E, Ioannidis JPA. Meta-analysis in genome-wide association studies. *Pharmacogenomics* 2009;**10**:191–201.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc* 1995;**57**:289–300.
17. Seminara D, Khoury MJ, O'Brien TR, *et al*. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* 2007;**18**:1–8.
18. Minelli C, Thompson JR, Abrams KR, *et al*. The quality of meta-analyses of genetic association studies: a review with recommendations. *Am J Epidemiol* 2009;**170**:1333–43.
19. Ioannidis JPA, Thomas G, Daly MJ. Validating augmenting and refining genome-wide association signals. *Nat Rev Gen* 2009;**10**:318–29.
20. Ioannidis JPA, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* 2007;**2**:e841.
21. Kraft P, Zeggini E, Ioannidis JPA. Replication in genome-wide association studies. *Stat Sci* 2009;**24**:561–73.
22. Greene CS, Penrod NM, Williams SM, *et al*. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One* 2009;**4**:e5639.
23. Anonymous. Freely associating. *Nat Genet* 1999;**22**:1–2.
24. Chanock SJ, Manolio T, Boehnke M, *et al*. Replicating genotype-phenotype associations. *Nature* 2007;**447**:655–60.
25. Skol AD, Scott LJ, Abecasis GR, *et al*. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;**38**: 209–13.
26. Thomas DC, Casey G, Conti DV, *et al*. Methodological issues in multistage genome-wide association studies. *Stat Sci* 2009;**4**:414–29.
27. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet* 2010;**86**:6–22.
28. Roeder K, Wasserman L. Genome-wide significance levels and weighted hypothesis testing. *Stat Sci* 2009;**24**: 398–413.
29. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. *Plos Genetics* 2005;**1**:e32.
30. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;**11**: 499–51.
31. Thakkinstian A, Thompson J, Minelli C, *et al*. Choosing between per-genotype per-allele and trend approaches for initial detection of gene-disease association. *J Appl Statist* 2009;**36**:633–46.
32. Zheng G, Joo J, Zaykin D, *et al*. Robust tests in genome-wide scans under incomplete linkage disequilibrium. *Stat Sci* 2009;**24**:503–16.
33. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet Epidemiol* 2010;**34**:60–66.

34. Karp DR, Carlin S, Cook-Deegan R, *et al*. Ethical and practical issues associated with aggregating databases. *PLoS Med* 2008;**5**:e190.
35. Pfeiffer R, Gail M, Pee D. On combining data from genome-wide association studies to discover disease-associated SNPs. *Stat Sci* 2009;**24**:547–56.
36. Lebrec J, Stijnen T, van Houwelingen H. Dealing with heterogeneity between cohorts in genomewide SNP association studies. *Stat Appl Genet Mol Biol* 2010;**9**:1–20.
37. Lambert P, Sutton A, Abrams K, Jones D. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol* 2002; **55**:86–96.
38. Berlin J, Santanna J, Schmid C, *et al*. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statist Med* 2002;**21**:371–87.
39. McKinney BA, Reif DM, Ritchie MD, *et al*. Machine learning for detecting gene-gene interations: a review. *Appl Bioinform* 2006;**5**:77–88.
40. Trégouët D-A, König IR, Erdmann J, *et al*. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet* 2009;**41**:283–5.
41. Ioannidis JP, Ntzani EE, Trikalinos TA, *et al*. Replication validity of genetic association studies. *Nat Genet* 2001;**29**: 306–9.
42. Clarke GM, Carter KW, Palmer LK, *et al*. Fine mapping versus replication in while-genome association studies. *Am J Hum Genet* 2007;**81**:995–1005.
43. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Contr Clin Trials* 1986;**7**:177–88.
44. Greenland S. Invited commentary: A critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**: 290–6.
45. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Statist Soc A* 2009;**172**:137–59.
46. Storey JS, Tibshirani R. Statistical significance for genome-wide studies. *PNAS* 2003;**100**:9440–5.
47. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**: 1351–5.
48. Higgins JP, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.
49. Ioannidis JPA. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered* 2007;**64**: 203–13.
50. Lin P-I, Vance JM, Pericak-Vance MA, *et al*. No gene is an island: the flip-flop phenomenon. *Am J Hum Genet* 2007;**80**: 531–8.
51. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001;**358**:1356–60.
52. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003;**361**:598–604.
53. Clayton DG, Walker NM, Smyth DJ, *et al*. Population structure differential bias and genomic control in a large-scale case-control association study. *Nat Genet* 2005; **37**:1243–6.
54. Price AL, Patterson NJ, Plenge RM, *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–9.
55. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;**273**:1516–17.
56. Clayton D. P-values false discovery rates and Bayes factors: how should we assess the 'significance' of genetic associations? *Ann Hum Gen* 2003;**67**:630.
57. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008;**32**:227–34.
58. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998;**316**:1236–8.
59. Rice TK, Schork NJ, Rao DC. Methods for handling multiple testing. *Adv Genet* 2008;**60**:293–308.
60. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 2007;**80**:605–15.
61. Ghosh A, Zou F, Wright F A. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am J Hum Genet* 2008;**82**:1064–74.
62. Bowden J, Dudbridge F. Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genet Epidemiol* 2009;**33**:406–18.
63. Zhong H, Prentice RL. Correcting 'winner's curse' in odds ratios from genome-wide association findings for major complex human diseases. *Genet Epidemiol* 2010;**34**:78–91.
64. Wellcome Trust Case-Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature* 2007;**447**:661–78.
65. Samani NJ, Erdmann J, Hall AS, *et al*. Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007; **357**:443–53.
66. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 2009;**10**:681–90.
67. Wakefield J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet Epidemiol* 2009;**33**: 79–86.
68. Afilalo J, Duque G, Steele R, *et al*. Statins for secondary prevention in elderly patients A hierarchical Bayesian meta-analysis. *J Am Coll Cardiol* 2008;**51**:37–45.
69. Berry D, Wathen JK, Newell M. Bayesian model averaging in meta-analysis: vitamin E supplementation and mortality. *Clin Trials* 2009;628–41.
70. Conlon EM, Postier BL, Methe BA, *et al*. Hierarchical Bayesian meta-analysis models for cross-platform microarray studies. *J Appl Stat* 2009;**36**:1067–85.
71. De la Cruz O, Wen X, Ke B, *et al*. Gene region and pathway level analyses in whole-genome studies. *Genet Epidemiol* 2010;**34**:222–31.
72. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008;**299**:1335–44.
73. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;**55**:997–1004.