Review TRENDS in Genetics Vol.20 No.9 September 2004

Meta-analysis of genetic association studies

Marcus R. Munafò and Jonathan Flint

Cancer Research UK GPRG, Department of Clinical Pharmacology, Radcliffe Infirmary, University of Oxford, Oxford OX2 6HE, UK

Meta-analysis, a statistical tool for combining results across studies, is becoming popular as a method for resolving discrepancies in genetic association studies. Persistent difficulties in obtaining robust, replicable results in genetic association studies are almost certainly because genetic effects are small, requiring studies with many thousands of subjects to be detected. In this article, we describe how meta-analysis works and consider whether it will solve the problem of underpowered studies or whether it is another affliction visited by statisticians on geneticists. We show that meta-analysis has been successful in revealing unexpected sources of heterogeneity, such as publication bias. If heterogeneity is adequately recognized and taken into account, meta-analysis can confirm the involvement of a genetic variant, but it is not a substitute for an adequately powered primary study.

A quick search through PubMed should be enough to convince anyone that the annual number of published (and presumably peer-reviewed) genetic association studies is increasing exponentially. We estimate that, in the field of psychiatric genetics alone, the rate is currently about one paper per day. It is also abundantly clear that the volume of reports is no index of the reliability of the results: each convincing association can be paired with an equally convincing rebuttal, followed in turn by another positive finding, seemingly *ad infinitum*. A survey of 600 positive associations between common gene variants and disease showed that most reported associations are not robust: of 166 associations that were studied three or more times, only six were replicated consistently [1]. What is going on?

Once we agree that physicians do really know how to measure blood pressure and diagnose diabetes, that psychiatrists can identify mental illness and that the rate of genotyping error does not completely invalidate the study (in other words that our measurement of the dependent and independent variables is reliable), then the remaining culprit is the design of genetic association studies. Delightfully simple in principle (just compare the allele frequencies in a selection of cases and controls and look for a statistically significant difference), it has nevertheless provided statistical geneticists fodder for almost as many journal pages as the association studies themselves. For some time POPULATION STRATIFICATION (see Glossary) has been blamed [2]; it became de rigueur to use the TRANSMISSION DISEQUILIBRIUM TEST to ensure the publication of a genetic association test or, more cunningly, to employ a GENOMIC CONTROL (more journal space is now being devoted to Monte Carlo Markov Chains and the like) [3]. More recently, examining haplotype structure and, inevitably, developing novel statistical methods to employ haplotypes in association tests has been in vogue [4,5]. Now, a (relatively) new solution to the problem of inconsistent findings is to use meta-analysis. Will it help?

At least it is proving to be popular. In 1984, Green and Hall described the potential value of meta-analytic investigation [6]. In that year there were 34 English-language citations in Medline that included the key word 'meta-analysis' and 89 citations in PsychInfo–PsychLit. By 1999 the corresponding number of such citations was 823 and 262, respectively. A similar pattern exists for the meta-analyses of genetic association studies: between 1994 and 1998 there were 27 published meta-analyses of genetic association studies, whereas between 1999 and

Glossary

Population stratification: occurs when a population consists of a set of subpopulations. If one subpopulation contains a frequency of disease allele that is relatively high, then any marker also at a higher frequency will appear to be associated, wherever it is located in the genome.
Transmission disequilibrium test: a method of detecting genetic association that avoids problems of population stratification. Instead of comparing unrelated cases and controls the test determines whether, given the parental genotypes, the alleles that are transmitted from parent to child and the child's affectation status are independent.
Genomic control: a method to assess population stratification by using data from a series of unlinked markers.
Relative risk: relative risk is the ratio of the incidence of the phenotype under consideration in subjects with the variant allele to the incidence in those without the variant allele.
Odds ratio: this is closely related to the relative risk and is defined as the odds of possessing the phenotype in those with the variant allele divided by the odds of possessing the phenotype in those without the variant allele. Odds ratios are simply a different way of expressing this association than relative risk because they compare odds rather than risk of an event.
Type I error: the erroneous rejection of a true hypothesis (i.e. falsely rejecting the null hypothesis and thereby concluding that association exists).
Confounding: the failure to separate two variables. Therefore, their independent effects cannot be independently ascertained.
Z-score: the standardized expression of a value in terms of its relative position in the full distribution of values, relative to the mean of the distribution in standard deviation units. It therefore can be used to calculate a corresponding *P*-value (and vice versa).
Power: a measure of the probability that any given statistical test will detect a significant relationship when one actually exists in the data.
Effect-size estimates: tests of a null hypothesis can show that an effect is significant but not how large the effect is. Measures of effect size are based on the proportion of variance in the data that can be attributed to the experimental variables.

Corresponding author: Marcus R. Munafò (marcus.munafo@clinpharm.ox.ac.uk).
Available online 17 July 2004

 0168-9525/$ - see front matter doi:10.1016/j.tig.2004.06.014

2003 there were 90 such meta-analyses (Figure 1). The reason for the popularity of this method is that it has the potential to deal with a major shortcoming in most genetic association studies: lack of POWER.

Questions of power

Where the effect sizes are large, even rare alleles (such as those giving rise to mental retardation) have been successfully identified [7]. For many linkage and association studies the failure to provide convincing evidence of linkage, even in large sample sizes, indicates that the RELATIVE RISK of disease susceptibility loci is moderate to low. Even results from successfully replicated association studies indicate that the relative risk attributable to a single locus is small [8]. For example, a T-cell regulatory gene [cytotoxic T-lymphocyte-associated protein 4 (*CTLA4*)], identified as a susceptibility locus for autoimmune disease, has an ODDS RATIO of 1.5 [9]. A relative risk of 1.5 has been suggested as reasonable for a gene conferring susceptibility to breast cancer [10].

Taken together, the data from both linkage and association studies indicate that susceptibility loci for common disease have a small effect. Simply put, this means that genetic association studies need to test thousands of cases and controls to have a reasonable chance of finding an effect. For an odds ratio of 1.3 (probably a typical value), Zondervan and Cardon show that in the best cases, where the marker and disease allele frequencies match, sample sizes of 2000–10 000 cases and controls are required to obtain 80% power [11].

Although a single laboratory might not be able to obtain such numbers, the combined world literature might if there was a way to analyse the data jointly. Meta-analysis goes some way to providing a tool to do just that. Although the basic elements of meta-analytic techniques can be traced back to Fisher [12], the term 'meta-analysis' was coined by Glass in 1976 [13] and refers to the synthesis of disparate datasets to ascertain a summary conclusion that is derived from the global body of data. It is a quantitative approach to combine systematically the results of previous research to arrive at a conclusion about a body of evidence.

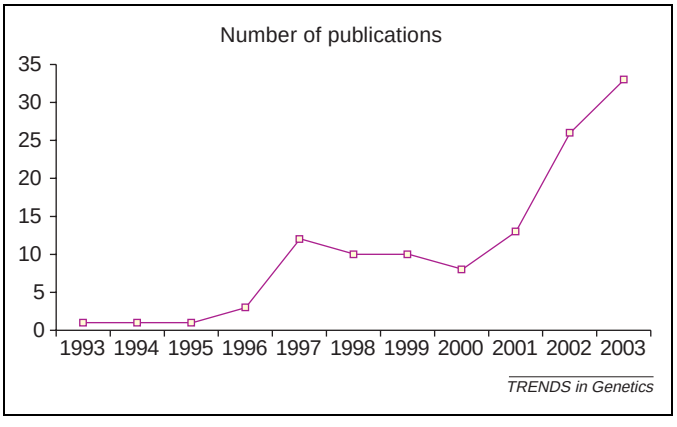

Figure 1. Publication trends for meta-analysis of genetic association studies. The number of publications in Medline and PsychInfo that included the key words 'meta-analysis' and 'genetic association' in the past decade. Used appropriately, meta-analysis has the potential to overcome the lack of power that is a problem in so many genetic association studies.

Consider the evidence supporting the role of calpain 10 (*CAPN10*) in type 2 diabetes susceptibility [14]. After the initial report that this gene conferred a threefold increased in risk in a population of Mexican Americans, there followed a fairly typical series of conflicting studies, using different populations and different markers. Collecting all case-control studies yielded 3303 subjects that were suitable for meta-analysis [15]. Of the ten studies, only one reported a significant association at $P<0.05$ [16], but the meta-analysis, rather than favouring the results of the majority of studies that produced no association, yielded a P-value of 0.02. The authors then performed a further large-scale association study of one single nucleotide polymorphism in *CAPN10* and reported an odds ratio of 1.17 with a P-value of 0.007.

Meta-analysis methodology

How is this 'magic' achieved? The method does not require the use of the raw genotypes (which would be ideal, although it is important to include the study contributing each set of raw data as a covariate) but instead calculates the effect size that each study attributes to the genetic variant, weighted according to the study size; small studies contribute less than large studies because they are likely to give less accurate EFFECT-SIZE ESTIMATES. Meta-analysis uses two types of effect-size estimate: differences between group means (effect size d) or zero-order correlations (effect size r). The zero-order correlations can be calculated from χ^2 and z-SCORE statistics and from a simple P-value [17].

One way of examining how meta-analysis functions is to consider what happens when we combine the P-values from several studies. Traditionally, the statistical sciences have relied on significance testing to evaluate the importance of reported data. This has the potential to mislead because researchers are often falsely seduced by a single highly significant report, while failing to appreciate that several reports that indicate comparable effect-size estimates (even if all of these, individually, fail to reach statistical significance) provide stronger support for this effect. For example, two results (from homogeneous studies, presumably) at $P=0.06$ provide substantially stronger evidence ($P=0.014$) against the null hypothesis than a single result at $P=0.05$. Similarly, ten results at $P=0.10$ provide stronger evidence ($P=0.000025$) against the null hypothesis than five results at $P=0.05$ ($P=0.00012$) [17].

If we have several studies measuring a continuous variable in two genotype groups the statistic that would be used is an estimate of d (i.e. a difference between group means), calculated by dividing the difference in the two group means within each study by the standard deviation (SD) difference (typically estimated from the individual group SDs). This would generate a standardized estimate of d for each study, which could then be combined by weighting each estimate by the sample size of the study that provided it.

It is also possible to combine P-values directly, by converting these P-values to z-scores and calculating these using the following formula, where k is the number of individual studies contributing to the meta-analysis

(i.e. the meta-analytic analogue of n):

$$Z_{\text{overall}} = \sum Z_i/\sqrt{k} \quad \text{(Eqn 1)}$$

This is then followed by the conversion of the summary estimate z-score back to a P-value. This is a relatively crude technique and has the disadvantage of not providing an effect-size estimate. Surprisingly, the increased power afforded by a meta-analysis does not simply result from the aggregation of raw data across studies and the subsequent large sample that is generated. A lot depends on the way the data are combined.

Two models can be used to combine the individual effect-size estimates (e.g. the estimate of the contribution of a postulated allele to the phenotype) provided by primary studies: a fixed-effects model or a random-effects model. Fixed-effects analysis assumes that all study samples are derived from a single population with a common effect size [18]. Therefore, within a fixed-effects model, only sampling error (theoretically) contributes to the differences between the observed effect-size estimates across individual studies. By contrast, random-effects analysis assumes that the study samples included in a meta-analysis can be drawn from a distribution of populations (so that there might be subpopulations for which there is no effect and others for which there is a substantial effect). Within a random-effects model, two sources of variance exist that contribute to the observed differences between effect-size estimates: sampling error and between-study heterogeneity. Between-study heterogeneity can be due to any potentially relevant differences between the study designs and methodologies, such as populations from which the study samples are drawn. For example, in genetic association studies, the causes of between-study heterogeneity include: (i) the possibility that an association exists in one population but not in another; (ii) the possibility that different studies did not use comparable measures of phenotype; or (iii) the possibility that allelic distributions deviated from Hardy–Weinberg equilibrium in some studies.

Fixed-effects- and random-effects analyses address fundamentally different research questions. The former asks what the best estimate of the true effect size is in the population studied, whereas the latter asks what the range and distribution of effect sizes is in the sample of populations studied. Therefore, the calculation of the mean of the distribution of population effect sizes (random-effects model) provides different information from the calculation of the mean of the distribution of sample effect sizes (fixed-effects model).

In fixed-effects analysis, increasing the number of studies contributing to the analysis will result in an increase in power because additional studies will result in a narrowing of the confidence intervals around the effect-size estimate. There is a marked risk of TYPE I ERROR, however, if there is substantial between-study heterogeneity. By contrast, increasing the number of studies contributing to a random-effects analysis will not necessarily result in an increase in power because of the possibility that the addition of studies will also result in the addition of larger variance-component estimates, if the addition of studies increases the total between-study heterogeneity [18].

Dealing with heterogeneity

A crucial question for any meta-analysis is the degree of heterogeneity that exists between the individual studies. Indeed, it has been argued that meta-analysis is analogous to averaging the characteristics of apples and oranges [19] and, consequently, its outcome is meaningless. Heterogeneity can be identified graphically by examining the SD of the effects sizes for each contributing study and looking for outliers and clusters, or by using a χ^2 test of heterogeneity. The χ^2 test is potentially flawed because a significant χ^2 test might be obtained in a large sample but not in a small sample, even if the degree of heterogeneity present in each case is the same. It is advisable, therefore, to use a combination of formal statistical and graphical methods to assess the degree and sources of between-study heterogeneity. For example, clusters of studies identified graphically might share common characteristics (e.g. ethnicity) that can be an important moderator of any genetic association.

Heterogeneity between studies is, perhaps not surprisingly, common. In 2001, Ioannidis *et al.* conducted a meta-analysis of 370 studies addressing 36 genetic associations [20]. They found that significant between-study heterogeneity is frequent, and that the results of the first study often correlate only modestly with subsequent research on the same association. One might think that heterogeneity can be taken into account by applying a random-effects model; but this can introduce an unintentional statistical 'sleight of hand' [18] because fixed-effects- and random-effects models address different research questions (see previous discussion). A more effective way to deal with heterogeneity is to identify its causes and then incorporate them directly in the statistical model that tests for genetic association.

Sensitivity analysis and publication bias

Investigating heterogeneity can also give rise to findings beyond the scope of an individual genetic association study. For example, in a recent meta-analysis of the genetic effects on personality, we found evidence of heterogeneity that, when a random-effects model was used, indicated that there was no evidence for an association with any of the candidate genes [21]. We combined studies that used different measures of personality because factor analytic and correlation techniques have shown substantial equivalence between at least two personality dimensions [22–24]. These phenotypes, extraversion and neuroticism, can be measured by several different questionnaire instruments that are frequently assumed to be broadly comparable [25–28]. Nevertheless, we found evidence for heterogeneity as a result of the questionnaire used, an unexpected conclusion that was also reported by another meta-analysis of personality studies [29]. This suggests that one questionnaire instrument can provide a stronger genetic signal than another.

Subgroup analysis (sometimes known as sensitivity analysis) can also be used to assess the impact of heterogeneity. Performing meta-analysis both with and

without potentially problematic studies is a simple method of assessing the impact of a potential source of heterogeneity. These 'problem' studies can be identified as clusters or outliers using graphical methods, and the case for their exclusion will be strengthened if they share some unique characteristic (such as allele frequencies that are not in Hardy–Weinberg equilibrium). There can also be *a priori* reasons for excluding certain studies on the grounds of quality control (because, for example, departure from Hardy–Weinberg equilibrium can indicate the presence of genotyping error). An important example of the value of sensitivity analysis is dealing with publication bias. Publication bias can exist when non-significant findings remain unpublished, thereby artificially inflating the apparent magnitude of an effect. The concern is not new and was raised almost 50 years ago in relation to psychiatric and psychological research [30]. Tests of publication bias can be divided into graphical tests and formal statistical tests.

The funnel plot is a commonly used graphical test that assesses publication bias in meta-analytic datasets [31]. The rationale behind this test is that if all studies come from a single population then the plot should look like a funnel with the diameter of the funnel decreasing (i.e. effect-size estimate becoming more accurate) as the sample size increases [32]. In the absence of publication bias, one should expect a symmetrical funnel plot; asymmetry is therefore suggestive of the possibility of publication bias. An alternative graphical test of publication bias can be derived by assessing the linearity of the normal quantile plot [32]. This scatter plot compares the quantiles of an observed sample distribution with the quantiles of the standard normal distribution. If the data from the observed sample are from a normal population, the resulting points should form a straight line. Any deviation from this (i.e. a curvilinear association) indicates the possibility of publication bias. Finally, a range of statistical tests also exist that enable the hypothesis that a publication bias exists to be tested formally, without the subjectivity that is inherent in a visual inspection of a graphical test.

Rosenthal called publication bias the 'file drawer problem', and proposed that it could be assessed by calculating the 'fail-safe N': the number of hypothetical (i.e. potentially unpublished) negative studies (i.e. studies in which the effect size is zero) that would be needed to increase the *P*-value for the meta-analysis to >0.05 [17]. Several alternative approaches are based on the assumption that the results from an individual study affect its probability of publication. These methods are called 'selection models' [33–35] and can be extended to estimate subgroup effects [36], corrected for the estimated publication bias.

Multivariate analysis

Usually meta-analysis relies on the assessment of main effects, with subgroup analyses being used to investigate different levels of potential moderator variables, as described previously. An alternative approach is to perform a multivariate meta-analysis, in the form of a meta-regression, with the inclusion of covariates within this framework. This approach enables the moderating effect of a covariate, such as sex or ethnicity, to be tested formally [37]. In contrast to simple meta-analysis, meta-regression relates effect size to one or more characteristics of the studies included [38], and the potential utility of this approach has resulted in a marked increase in the use of meta-regression in meta-analytic reviews. Such covariates can be derived from the participant characteristics of the included studies. For example, if sex is to be included as a covariate in a meta-regression framework the percentage of female participants in each study can be used as the covariate. Meta-regression can occur within a fixed-effects- and a random-effects framework.

Meta-regression investigates whether particular covariates, or effect modifiers, explain any of the observed between-study heterogeneity effects. Therefore, it is usually appropriate to use a random-effects model because this accommodates the possibility that the underlying effect differs across studies. It is appropriate to use meta-regression even when graphical and formal statistical tests do not indicate heterogeneity, particularly if there are grounds for believing a potential moderator variable is important [37]. Indeed, simply performing multiple analyses on each potential covariate for which data exists is potentially hazardous and carries a high risk of Type I error. Pre-specifying which covariates are of interest, and limiting the number to be investigated, can protect against this danger.

Considering subgroup analysis formally as a meta-regression has several advantages because it focuses on the differences between subgroups, rather than considering the effect in each subgroup separately. Moreover, in the context of a random-effects model, allowance is made for residual heterogeneity that is not accounted for by the subgrouping. It is important to bear in mind, however, that meta-regression, although enabling covariates to be considered, does not have the methodological rigour of a properly designed study that is intended to test the effect of these covariates formally. Therefore, meta-regression can potentially suffer from bias by CONFOUNDING. One common limitation is the number of studies that are available for inclusion. With a small number of data points (i.e. individual studies) the model might become unstable and the potential for drawing robust conclusions limited.

What to include and how to determine heterogeneity?

A meta-analysis is only as good as its constituent parts and its success is determined in part by the diligence of the investigators identifying suitable studies and extracting the correct information. Lohmueller and colleagues analysed 301 studies that tested 25 genetic associations [8]. They found there was a large excess of studies replicating the first positive reports (evidence against no true association) and that almost a third of the associations (8/25) were replicated in a meta-analysis of follow-up studies. If we compare these results with another large-scale survey, we find a good degree of agreement. Ioannidis and colleagues tested 55 meta-analyses and found that in nine (16%) the genetic association was

Box 1. Recommendations for the use of meta-analysis

By increasing the ability to detect small effects, meta-analysis has the potential to solve a persistent problem in human genetic association studies. For many conditions, the literature is replete with small studies reporting inconsistent findings. As has been recently pointed out, there are no standards for designing, implementing or reporting genetic association studies [40]. We suggest that meta-analysis can determine whether a real effect is present or not and we recommend its use to address this issue; at the same time, we hope that the application of meta-analysis does not suffer from the same lack of standardization that has plagued individual genetic association studies.

Therefore, we suggest that the meta-analysis should include sufficient information for the results to be assessed for completeness, quality control and robustness. First, the steps taken to ascertain studies for analysis and the criteria for inclusion of subjects must be stated. Small differences in selection criteria can alter fundamentally the outcome of the meta-analysis; therefore, the selection procedure must be described in detail. We recommend that either quality control is included as a covariate in the analysis so that the importance of ascertainment can be tested, or subgroup analyses are conducted excluding studies of potentially low quality.

Second, the meta-analysis should assess all of the included studies for the presence of heterogeneity and should attempt to determine the possible causes of heterogeneity, such as ethnicity, gender and phenotype. In addition, there are likely to be other, phenotypic-specific effects that should be tested for heterogeneity. Note that successful large-scale studies reporting significant effects do not rule out the existence of between-study heterogeneity; performing meta-analysis on a group of studies, even if all of them report a significant result for the same genetic variant, is still worthwhile.

Third, the genotypes of all component studies should be tested for Hardy–Weinberg equilibrium and excluded if there is a significant deviation. Remarkably, not all studies include this quality control check on their data.

Finally, evidence of publication bias should be sought and reported. Any conclusions obtained from a meta-analysis in which publication bias might be suspected must be tempered, and the possible effect (either to increase or decrease the likelihood of finding a genetic association) should be noted.

replicated without heterogeneity. Both groups agree that effects sizes are small (mean odds ratios 1.33, SD 0.65 [39]). Although they disagree on the degree of replication (8/25 compared with 9/55), the difference is not significant. However, as Ioannidis *et al.* point out, 'to disentangle bias from true heterogeneity may often be difficult, since there is no gold standard test for discrimination'.

Concluding remarks

Meta-analysis is a potentially powerful tool for assessing population-wide effects of candidate genes on complex phenotypes (Box 1) and can provide evidence of previously unexpected diversity, for example, by revealing heterogeneity in studies of apparently similar populations [20,39]. But the results are only as good as the input data, and the reporting of this data can constrain the extent to which a meta-analysis might be informative or even performed. In genetic association studies, this can often mean, for example, being constrained to investigate two genotype groups only because smaller studies frequently combine rare homozygote and heterozygote groups to increase their statistical power (e.g. homozygous wild-type groups compared with combined heterozygous and homozygous mutant groups). Meta-analysis is not a replacement for adequately powered genetic association studies. Perhaps the greatest value of a meta-analysis is to tell us what we need to aim for: the small odds ratios (a mean of 1.33 in 55 meta-analyses [39]) indicate that studies will need to include many thousands of subjects if they are to provide unequivocal evidence of an association between a genetic variant and a phenotype [11].

Acknowledgements

M.M. holds a Cancer Research UK research fellowship. J.F. is supported by the Wellcome Trust.

References

1 Hirschhorn, J.N. *et al.* (2002) A comprehensive review of genetic association studies. *Genet. Med.* 4, 45–61
2 Cardon, L.R. and Palmer, L.J. (2003) Population stratification and spurious allelic association. *Lancet* 361, 598–604
3 Pritchard, J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181
4 Wall, J.D. and Pritchard, J.K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 4, 587–597
5 Goldstein, D.B. *et al.* (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* 19, 615–622
6 Green, B. and Hall, J. (1984) Quantitative methods for literature review. *Annu. Rev. Psychol.* 35, 37–53
7 Chelly, J. and Mandel, J.L. (2001) Monogenic causes of X-linked mental retardation. *Nat. Rev. Genet.* 2, 669–680
8 Lohmueller, K.E. *et al.* (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33, 177–182
9 Ueda, H. *et al.* (2003) Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. *Nature* 423, 506–511
10 Pharoah, P.D. *et al.* (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* 31, 33–36
11 Zondervan, K.T. and Cardon, L.R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* 5, 89–100
12 Fisher, R.A. (1925) *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh
13 Glass, G.V. (1976) Primary, secondary and meta-analysis. *Educ. Res.* 5, 3–8
14 Horikawa, Y. *et al.* (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* 26, 163–175
15 Weedon, M.N. *et al.* (2003) Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am. J. Hum. Genet.* 73, 1208–1212
16 Evans, J.C. *et al.* (2001) Studies of association between the gene for calpain-10 and type 2 diabetes mellitus in the United Kingdom. *Am. J. Hum. Genet.* 69, 544–552
17 Rosenthal, R. and DiMatteo, M.R. (2001) Meta-analysis: recent developments in quantitative methods for literature reviews. *Annu. Rev. Psychol.* 52, 59–82
18 Cohn, L.D. and Becker, B.J. (2003) How meta-analysis increases statistical power. *Psychol. Methods* 8, 243–253
19 Hunt, M. (1997) *How Science Takes Stock: The Story of Meta-Analysis*, Russell Sage Foundation, New York, USA
20 Ioannidis, J.P. *et al.* (2001) Replication validity of genetic association studies. *Nat. Genet.* 29, 306–309
21 Munafo, M.R. *et al.* (2003) Genetic polymorphisms and personality in healthy adults: a systematic review and meta-analysis. *Mol. Psychiatry* 8, 471–484
22 Costa, P.T. and McCrae, R.R. (1992) *Revised NEO Personality*

Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual, Psychological Assessment Resources, Lutz, Florida, USA
23 Zuckerman, M. *et al*. (1993) Comparison of three structural models for personality: the big three, the big five, and the alternative five. *J. Pers. Soc. Psychol.* 65, 757–768
24 Aluja, A. *et al*. (2002) A comparative study of Zuckerman's three structural models for personality through the NEO-PI-R, ZKPQ-III-R, EPQ-RS and Goldberg's 50-bipolar adjectives. *Pers. Individ. Dif.* 33, 713–725
25 Eysenck, H. and Eysenck, M.W. (1985) *Personality and Individual Differences. A Natural Science Approach*, Plenum Press, London
26 Costa, P.T. and McCrae, R.R. (1985) *The NEO Personality Manual*, Psychological Assessment Resources, Lutz, Florida, USA
27 Caprara, G.V. *et al*. (1993) *BFQ – Big Five Questionnaire: Manual*, Organizzioni Speciali, Florence, Italy
28 Zuckerman, M. *et al*. (1991) Five (or three) robust questionnaire scale factors of personality without culture. *Pers. Individ. Dif.* 12, 929–941
29 Sen, S. *et al*. (2004) Meta-analysis of the association between a serotonin transporter promoter polymorphism (5-HTTLPR) and anxiety-related personality traits. *Am. J. Med. Genet.* 127B, 85–89
30 Sterling, T.D. (1959) Publication decision and their possible effects on inferences drawn from tests of significance – or vice versa. *J. Am. Stat. Assoc.* 54, 30–34
31 Egger, M. and Davey-Smith, G. (1997) Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315, 629–634
32 Wang, M.C. and Bushman, B.J. (1998) Using normal quantile plots to explore meta-analytic data sets. *Psychol. Methods* 3, 46–54
33 Dear, K.B.G. and Begg, C.B. (1992) An approach to assessing publication bias prior to performing a meta-analysis. *Stat. Sci.* 7, 237–245
34 Hedges, L.V. (1992) Modeling publication selection effects in meta-analysis. *Stat. Sci.* 7, 246–255
35 Iyengar, S. and Greenhouse, J.B. (1988) Selection problems and the file drawer problem. *Stat. Sci.* 3, 109–135
36 Vevea, J.L. and Hedges, L.V. (1995) A general linear model for estimating effect size in the presence of publication bias. *Psychmetrika* 60, 419–435
37 Munafo, M.R. *et al.* Are there sex differences in the association between the *5HTT* gene and neuroticism? A meta-analysis. *Pers. Individ. Dif.* (in press)
38 Thompson, S.G. and Higgins, J.P.T. (2002) How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* 21, 1559–1573
39 Ioannidis, J.P. *et al*. (2003) Genetic associations in large versus small studies: an empirical assessment. *Lancet* 361, 567–571
40 Becker, K.G. *et al*. (2004) The genetic association database. *Nat. Genet.* 36, 431–432