Replication validity of genetic association studies

John P.A. Ioannidis[1–3], Evangelia E. Ntzani[1], Thomas A. Trikalinos[1] & Despina G. Contopoulos-Ioannidis[1,4]

Published online: 15 October 2001, DOI: 10.1038/ng749

The rapid growth of human genetics creates countless opportunities for studies of disease association. Given the number of potentially identifiable genetic markers and the multitude of clinical outcomes to which these may be linked, the testing and validation of statistical hypotheses in genetic epidemiology is a task of unprecedented scale[1,2]. Meta-analysis provides a quantitative approach for combining the results of various studies on the same topic, and for estimating and explaining their diversity[3,4]. Here, we have evaluated by meta-analysis 370 studies addressing 36 genetic associations for various outcomes of disease. We show that significant between-study heterogeneity (diversity) is frequent, and that the results of the first study correlate only modestly with subsequent research on the same association. The first study often suggests a stronger genetic effect than is found by subsequent studies. Both bias and genuine population diversity might explain why early association studies tend to overestimate the disease protection or predisposition conferred by a genetic polymorphism. We conclude that a systematic meta-analytic approach may assist in estimating population-wide effects of genetic risk factors in human disease.

For each genetic disease association that we examined, a median of 9 studies (interquartile range 5–15) had been published. The main comparisons of disease cases and controls were based on allele frequencies (n=13), genotypes assuming recessive inheritance (n=16) or various other contrasts of genotypes (n=7; Table 1). In 14 of 36 cases (39%), there was statistically significant heterogeneity[5] between the results of the various studies on the same topic. The odds of having statistically significant heterogeneity between the studies of the same topic is greater when more studies were carried out (odds ratio 1.15 per additional study, P=0.02). In the ten meta-analyses with less than six studies each, we did not detect statistically significant heterogeneity between the results of the included studies. By contrast, we did detect statistically significant heterogeneity between the combined studies in 7 of 9 meta-analyses with at least 15 included studies. The power of a meta-analysis to detect heterogeneity increases with additional studies; alternatively, publication bias against studies with 'negative' or discordant results may be less prominent in fields where more studies are eventually published.

When we compared the strength of the postulated genetic association (as conveyed by the odds ratio) determined in the first study or studies against that determined in subsequent research across the 36 topics, the correlation was modest (fixed-effects modeling[6], r=0.42, P=0.011; random-effects modeling[7–9], r=0.51, P=0.002; Fig. 1).The first study tended to give more impressive estimates of disease protection or predisposition than subsequent research. This occurred in 25 or 26 of 36 cases (P=0.029 or P=0.011) depending on the modeling (fixed or random effects, respectively). Regardless of the modeling, in eight cases the discrepancy between the first and subsequent studies was beyond what would occur by chance alone (P<0.05). In another two cases, the discrepancy was beyond chance only when determined by fixed-effects modeling (Fig. 1*a*).

Figure 2*a* presents the eight meta-analyses in which the results of the first study differed significantly beyond chance (P<0.05) from those of subsequent research by both fixed- and random-effects calculations. The typical situation is that a very strong association is proposed by the first study, which becomes gradually less prominent or even disappears as more data accumulate[10,11]. Such behavior may suggest a spurious finding that is not validated by subsequent research, an exaggerated finding that eventually finds its appropriate measure or a gene effect that is stronger in some subpopulations than in others.

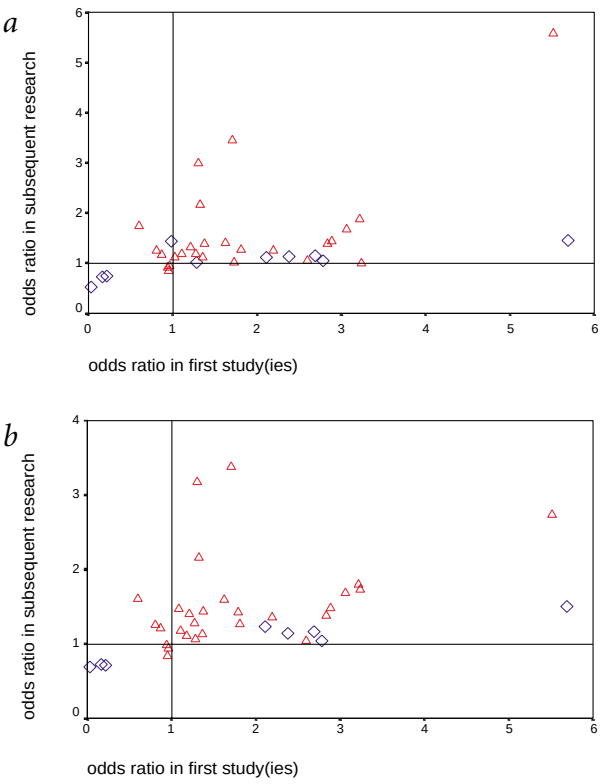

Fig. 1 Correlation between the odds ratio (OR) in the first study/studies and in subsequent research. OR>1 suggests predisposition towards the disease, whereas OR<1 suggests protection from the disease. Blue diamonds denote statistically significant discrepancies beyond chance between first and subsequent studies (***a***, fixed effects; ***b***, random effects).

[1]*Clinical and Molecular Epidemiology Unit and Clinical Trials and Evidence-Based Medicine Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece.* [2]*Ioannina Biomedical Research Institute, Ioannina 45110, Greece.* [3]*Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts 02111, USA.* [4]*Department of Pediatrics, George Washington University School of Medicine and Health Sciences, Washington DC 20010, USA. Correspondence and requests for materials should be addressed to J.I. (e-mail: jioannid@cc.uoi.gr).*

The first studies for these eight postulated associations appeared in prestigious journals (five in journals with an impact factor higher than 9.0; three in journals with an impact factor of 2.5–4.0), and invariably showed strong associations, with odds ratios as large as 2.1–5.7 (suggesting strong genetic predisposition) or as small as 0.03–0.22 (suggesting strong genetic protection). The subsequent studies targeting these same associations showed more marginal or no statistically significant effects at all. The subsequent studies were published in journals with lower average impact factors than the first studies in four topics, or with similar average impact factors as the first studies in another four topics.

Subsequent studies have failed to validate the originally proposed importance of dopamine receptor D3 gene polymorphisms for schizophrenia, of apolipoprotein E gene polymorphisms for dementia in patients with Down syndrome, of angiotensinogen gene polymorphisms for essential hypertension, of cytochrome p450 2D6 (*CYP2D6*) gene mutations for Parkinson disease or of *CYP2D6* metabolic status for lung cancer. Subsequent studies have confirmed that glutathione S-transferase M1 status may be important in susceptibility to lung cancer, that dopamine receptor D2 gene polymorphisms may confer some susceptibility to alcoholism and that angiotensin-converting enzyme gene polymorphisms may be involved in diabetic nephropathy; however, the strength of the associations found by the subsequent studies is significantly smaller than that postulated by the first studies for each of these three subjects.

Conversely, in eight other topics the first study or studies did not find a statistically significant difference between disease cases and controls but, with the accumulation of further data, the genetic association became formally statistically significant in the meta-analysis. The results of the first study or studies did not differ significantly from those of the subsequent research. Trajectories of these cumulative odds ratios are shown in Fig. 2*b*. For these eight topics, the first studies were published in journals with impact factors between 1.1 and 10.2.

Table 1 • Characteristics of the meta-analyses of 36 genetic disease associations

ID	Disease/outcome	Gene (polymorphism) – Genetic contrast[a]	Contrast type	Subjects[b] (studies)
1	MI	*ACE* (insertion/deletion) – DD *vs.* DI + II	genotype	18,664 (15)
2	IHD	*ACE* (insertion/deletion) – DD *vs.* DI + II	genotype	21,876 (17)
3	ICVD	*ACE* (insertion/deletion) – DD *vs.* DI + II	genotype	11,394 (6)
4	poor clozapine response	*HTR2A* (102T/C) – CC *vs.* CT+TT	genotype	733 (6)
5	poor clozapine response	*HTR2A* (H452Y) – YY *vs.* HY+HH	genotype	676 (5)
6	vascular disease	*MTHFR* (677C/T) – TT *vs.* CC	genotype[c]	6,947 (23)
7	lung cancer	*CYP2D6* (deficient oxidation) – poor metabolizers *vs.* others	genotype	5,162 (14)
8	dementia in Down syndrome	*APOE* (ε2/ε3/ε4) – allele ε2 *vs.* ε3+ε4	allele	1,130 (9)
9	schizophrenia	*DRD3* (*Bal*1) – 11+22 *vs.* 12	genotype[c]	5,121 (25)
10	bipolar affective disorder	*MAOA* (*Fnu*4HI) – allele 1 *vs.* 2	allele	962 (3)
11	bipolar affective disorder	*MAOA* (CA) – allele 122 *vs.* others	allele	1,932 (7)
12	bipolar affective disorder	*TH* (tetranucleotide repeat) – allele 1 *vs.* others	allele	2,901 (8)
13	unipolar affective disorder	*TH* (tetranucleotide repeat) – allele 1 *vs.* others	allele	1,128 (3)
14	NIDDM	*KCNJ11*/*KIR6.2-BIR* (E23K) – KK *vs.* EK+EE	genotype	888 (4)
15	lung cancer	*GSTM1* (gene deletion) – null/null *vs.* others	genotype	9,724 (21)
16	lung cancer	*CYP1A1* (4889A/G) – GG *vs.* AA+AG	genotype	2,392 (6)
17	lung cancer	*CYP1A1* (*Msp*I) – +/+ *vs.* others	genotype	4,263 (12)
18	MI	*SERPINE1*/*PAI1* promoter (4G/5G) – 4G/4G *vs.* 5G/5G	genotype[c]	1,910 (10)
19	Parkinson disease	*CYP2D6* (1934G/A) – allele 4 *vs.* others	allele	7,029 (14)
20	essential HTN	*AGT* (M235T) – allele T235 *vs.* M235	allele	4,698 (6)
21	cancer	*HRAS*/*HRAS1* (rare alleles) – rare *vs.* common alleles	allele	8,542 (24)
22	left ventricular hypertrophy	*ACE* (insertion/deletion) – allele D *vs.* I	allele	8,186 (12)
23	bladder cancer	*NAT2* (slow acetylation alleles) – slow/slow *vs.* others	genotype	5,836 (20)
24	ICVD	*APOE* (ε2/ε3/ε4) – allele ε4 *vs.* others	allele	3,632 (9)
25	nonsyndromic cleft lip	*TGFA* (*Taq*I) – allele 2 *vs.* 1	allele	5,272 (9)
26	alcoholism	*DRD2* (*Taq*IA) – allele A1 *vs.* A2	allele	3,826 (15)
27	ischemic stroke	*ACE* (insertion/deletion) – DD *vs.* DI + II	genotype	2,160 (6)
28	diabetic nephropathy	*ACE* (insertion/deletion) – II *vs.* ID+DD	genotype	5,393 (20)
29	NTD	*MTHFR* (677C/T) – TT *vs.* TC+CC	genotype	1,033 (4)
30	NTD (mother)	*MTHFR* (677C/T) – TT *vs.* TC+CC	genotype	1,160 (4)
31	NTD (father)	*MTHFR* (677C/T) – TT *vs.* TC+CC	genotype	815 (3)
32	IHD	*APOE* (ε2/ε3/ε4) – ε4/ε3+ε4/ε2+ε4/ε4 *vs.* ε3/ε3	genotype[c]	8,962 (9)
33	IHD	*LPL* (D9N) – ND *vs.* DD	genotype[c]	2,022 (3)
34	IHD	*LPL* (N291S) – SN *vs.* NN	genotype[c]	13,115 (4)
35	IHD	*LPL* (S447X) – XS *vs.* SS	genotype[c]	4,067 (5)
36	alcoholic liver disease	*CYP2E*/*CYP2E1* (*Rsa*I) – allele c2 *vs.* others	allele	4,178 (9)

[a]The various polymorphisms are named according to commonly used nomenclature. For ID 7, 15 and 23, some of the included studies inferred genotype from phenotype determinations. *ACE*, gene encoding for angiotensin-converting enzyme; *AGT*, gene encoding angiotensinogen; *APOE*, gene encoding apolipoprotein E; CVD, cerebrovascular disease; *CYP*, gene(s) encoding cytochrome P450; *DRD2*/*DRD3*, gene encoding dopamine receptor D2/D3; *GSTM1*, gene encoding glutathione-*S*-transferase M1; *HRAS*/*HRAS1*, gene encoding v-Ha-ras Harvey rat sarcoma viral oncogene homolog; HTN, hypertension; *HTR2A*, gene encoding 5-hydroxytryptamine receptor 2A; ICVD, ischemic cerebrovascular disease; IHD, ischemic heart disease; *KCNJ11*/*KIR6.2-BIR*, K^+ inwardly rectifying channel/β-cell inward rectifier, subfamily J, member 11; *LPL*, gene encoding lipoprotein lipase; *MAOA*, gene encoding monoamine oxidase A; MI, myocardial infarction; *MTHFR*, gene encoding methylenetetrahydrofolate reductase; *NAT2*, gene encoding *N*-acetyltransferase 2; NIDDM, non-insulin-dependent diabetes mellitus; NTD, neural tube defect; *SERPINE1*/*PAI1*, gene encoding serine proteinase inhibitor, clade E, member 1/plasminogen activator inhibitor 1; *TGFA*, gene encoding transforming growth factor A; *TH*, gene encoding tyrosine hydroxylase. [b]When the comparison is based on allele frequencies, then the number refers to the total number of typed alleles in the comparison rather than the number of subjects. [c]The comparison is based on a contrast of genotypes assuming a gene effect other than recessive.

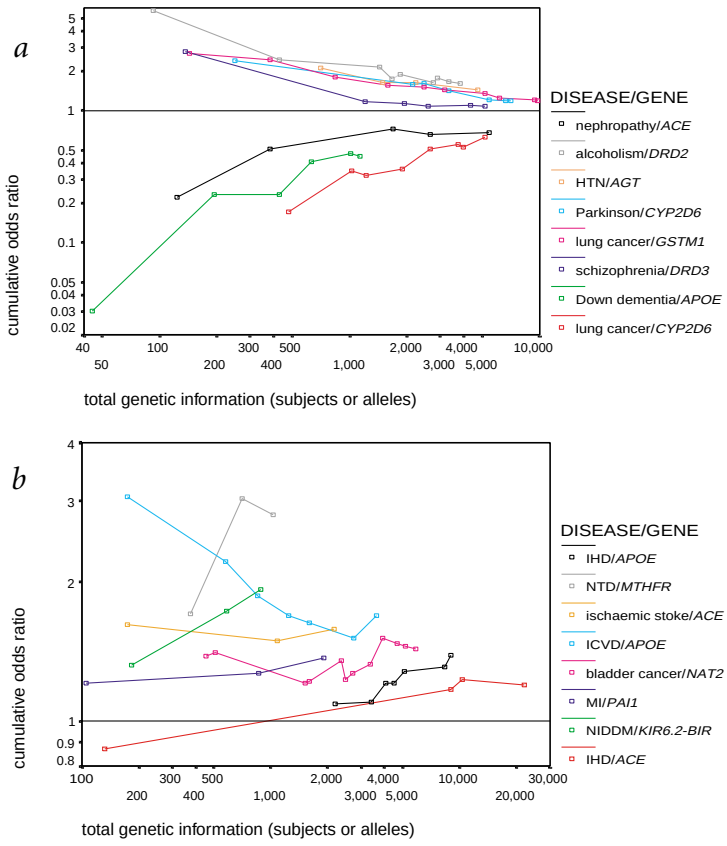

Fig. 2 Evolution of the strength of an association as more information is accumulated. The strength of the association is shown as an estimate of the odds ratio (OR) without confidence intervals. ***a***, Eight topics in which the results of the first study or studies differed beyond chance ($P<0.05$) when compared with the results of the subsequent studies. ***b***, Eight topics in which the first study or studies did not claim formal statistical significance for the genetic association but formal significance was reached by the end of the meta-analysis. Each trajectory starts at the OR of the first study or studies. Updated cumulative OR estimates are obtained at the end of each subsequent year, summarizing all information to that time (random effects). The horizontal axis (total genetic information) shows the total number of subjects genotyped with one of the contrasted genotypes, or the total number of typed alleles when specific allele frequencies are compared between disease cases and controls. Abbreviations are listed in Table 1.

In 12 other topics the first study or studies had not found statistical significance and this continued to be the case at the end of the meta-analysis. Finally, in the last eight subjects, the first study or studies had reached statistical significance and subsequent research did not disagree beyond chance; however, in only four of the eight was there formal statistical significance for the genetic association at the end of the meta-analysis.

Considering all 36 topics, the odds of finding a statistically significant discrepancy between the first and subsequent research is greater when more studies were carried out on the same issue, when the sample size of the first study or studies was smaller and when there was only a single first publication that had highlighted a clearly defined genetic contrast (Table 2). A small sample size of the first publication and a large number of studies were independent predictors of reaching discrepancies. We noted statistically significant discrepancies in 5 of 7 cases in which the first publications had a sample size of less than 150, compared with 3 of 29 when the sample size of the first study or studies was more than 150. Furthermore, we observed such discrepancies in 4 of 9 meta-analyses in which at least 15 studies had been published, whereas we documented no discrepancies among the 10 meta-analyses with 5 or fewer reports.

In summary, genetic association studies require cautious replication—an issue for both linkage and association studies. For linkage studies, other investigators have shown that replication is problematic under conditions of heterogeneity[12,13]. Pooling or Bayesian approaches with raw data may offer advantages[14–16], and data availability for such analyses would require coordinated efforts at an international level. Heterogeneity in the strength of an association is common even between studies of seemingly similar populations, which may differ in parameters that are not yet known or in parameters that the original studies have not captured. Meta-analysis may detect previously unrevealed diversity, and this should be pursued in subgroup analyses of raw data and in future studies. The evaluation of subgroup effects (such as racial differences or gene–environment interactions) is difficult[17] and requires large numbers of subjects[18].

We should also consider sampling biases[19]: given the large set of possible genetic associations probed by investigators worldwide, the most prominent findings represent an extreme sample and associations may be less extreme in new studies. Publication bias[20] and time-lag bias[21] are also possible: small studies with 'negative' statistically non-significant results may take longer to be published than 'positive' statistically significant studies[21]. Estimates of the size of a genetic effect may be inflated, if based only on a single study with impressive results. Finally, in some cases there may be large statistical uncertainty in the first study. Often genetic associations of disease are of modest magnitude (an odds ratio <2 or >0.5) and single studies are underpowered to detect them. Many studies of disease association are actually interested simply in gene detection rather than in estimating the size of the effect associated with a particular gene. Isolated statistical significance does not guarantee a genetic association, and lack of formal statistical significance does not exclude the possibility of an association.

Table 2 • Predictors of statistically significant discrepancies between the first and subsequent studies of the same genetic association

Predictor of discrepancy	Univariate regressions		Multivariate regression	
	OR (95% CI)	*P* value	OR (95% CI)	*P* value
total number of studies (per study)	1.17 (1.03–1.33)	0.020	1.18 (1.02–1.37)	0.028
sample size of first studies (doubling)	0.42 (0.17–0.98)	0.046	0.44 (0.19–0.99)	0.050
single first study with clear genetic contrast[a]	9.33 (1.01–86.3)	0.044	NS	NS

Comparisons are based on random-effects calculations. Odds ratios (OR) and 95% confidence intervals (CI) are derived from logistic regressions. The multivariate model is derived with backward elimination of variable according to likelihood ratio criteria. OR >1 suggests an increasing probability of finding significantly discrepant results beyond chance ($P<0.05$) between the first and subsequent studies. NS, not significant. [a]The single first study proposed a specific allele contrast or a specific genotype contrast to be the one reflecting the differential disease susceptibility.

Methods

Definitions. We considered meta-analyses of disease association studies that had been based on human genetic markers other than HLA alleles. A Medline search (last update December 2000) used the terms 'polymorphism(s)' and 'genetics', and 'meta-analysis' as type of publication. Meta-analyses with discrete outcomes qualified if (i) the outcome was related to a genetic marker (ii) detailed data were available for constructing 2×2 tables for each individual study (number of disease cases with and without the marker, number of controls with and without the marker) and (iii) data had been published in at least two different years. When a published report examined several genetic markers or several clinical outcomes, we considered separately each genetic marker and outcome. We obtained qualitatively similar results in analyses selecting only one outcome (the one with the largest number of tested subjects). We considered different polymorphisms of the same gene as separate genetic markers. When there were several meta-analyses on exactly the same association, we generally retained only the most updated one, provided it had adequate data for each of the included studies.

We screened in detail 51 reports of meta-analyses and excluded 25 of them (more recent similar meta-analysis available, n=19; lack of detailed data for individual studies, n=5; publication of all pertinent studies in the same year, n=1). Twenty-six meta-analysis reports were eligible (Web Notes A and B), addressing 36 genetic associations across 370 studies (Web Note C). When the genetic marker had more than two categories (such as AA homozygosity, Aa heterozygosity, aa homozygosity), we considered the comparison of the two categories that had been proposed originally in the first study in the field. One or both of the compared categories could be combinations of different genotypes (for example, Aa heterozygosity and aa homozygosity). When it was not clear which was the most important genetic contrast and when several 'first' studies were published in the same year in different journals, we selected the genetic contrast proposed by the meta-analysis. If several comparisons were carried out even by the meta-analysis, we used an *a priori* algorithm (which selects genotype contrasts over allele frequency comparisons, and genotype contrasts based on recessive inheritance over other genotype contrasts). For data presented only in abstract form and for unpublished data, we imputed publication to occur after the meta-analysis. We always considered the first study or studies for the specific genetic association, as well as all the other subsequent studies included in the meta-analysis. In 27 cases, there was an easily identifiable first study (Web Note B); in the other 9 cases, 2–10 studies had been published close to each other in the same year in different journals, and we therefore estimated their summary odds ratio by fixed- and random-effects models.

Modeling. We estimated between-study heterogeneity in each meta-analysis by the Q statistic; Q is traditionally considered to be significant for P<0.10 (ref. 5). Summary odds ratios calculations used the Mantel–Haenszel fixed-effects model[6] and the DerSimonian and Laird random-effects model[7,9]. Detailed graphs for each eligible meta-analysis are provided in the Supplementary Information (Web Figs. A1–A36). Fixed-effects models assume that all studies aim at evaluating a common truth and results differ by chance alone. Random-effects models anticipate that the studies may have genuine differences in their results[9]; thus, they also incorporate a between-study variance in their estimates. Random-effects models are generally more conservative (that is, they provide wider confidence intervals when there is between-study heterogeneity[5]). Fixed-effects models may be inappropriate if there is genuine heterogeneity in the size of genetic effects across subpopulations. Random-effects models are thus preferable, but they assume a certain distribution for the effect sizes that may be difficult to validate.

We estimated the Spearman rank-correlation coefficient between the odds ratios of first and subsequent studies across all the topics considered. We examined whether the results of the first published studies on a given association were significantly different beyond chance from the results of the subsequently published studies by using a z-score, as described previously[22]. We also constructed recursive cumulative meta-analysis plots that trace the evolution of the summary odds ratio as more data accumulate over time on the same topic[10,11]. Finally, we examined with logistic regressions[23] whether the number of studies, the sample size of the first study or studies and the availability of a single first study with a clearly proposed genetic contrast influenced the occurrence of statistically significant discrepancies between the results of the first and subsequent studies.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

This work was supported in part by a grant from the General Secretariat for Research and Technology, Greece, funded through the European Union.

Received 27 April; accepted 25 September 2001.

1. McCarthy, J.J. & Hilfiker, R. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nature Biotechnol.* **18**, 505–508 (2000).
2. Khoury, M.J. & Little, J. Human genome epidemiology reviews: the beginning of something HuGE. *Am. J. Epidemiol.* **151**, 2–3 (2000).
3. Lau, J., Ioannidis, J.P.A. & Schmid, C.H. Summing up evidence: one answer is not always enough. *Lancet* **351**, 123–127 (1998).
4. Mosteller, F. & Colditz, G.A. Understanding research synthesis (meta-analysis). *Annu. Rev. Public Health* **17**, 1–23 (1996).
5. Lau, J., Ioannidis, J.P. & Schmid, C.H. Quantitative synthesis in systematic reviews. *Ann. Intern. Med.* **127**, 820–826 (1997).
6. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst.* **22**, 719–748 (1959).
7. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
8. Hedges, L.V. & Olkin, I. *Statistical Methods for Meta-analysis* (Academic, Orlando, 1985).
9. Cooper, H. & Hedges, L.V. (eds.) *The Handbook of Research Synthesis* (Russell Sage Foundation, New York, 1994).
10. Ioannidis, J.P., Contopoulos-Ioannidis, D.G. & Lau, J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence for group and individual patient data. *J. Clin. Epidemiol.* **52**, 281–291 (1999).
11. Ioannidis, J.P. & Lau, J. Evolution of treatment effects over time: empirical insight from recursive cumulative meta-analyses. *Proc. Natl Acad. Sci. USA* **98**, 831–836 (2001).
12. Gershon, E.S. & Cloninger, C.R. (eds) *Genetic Approaches to Mental Disorders* (American Psychiatric Press, Washington, DC, 1994).
13. Vieland, V.J., Wang, K. & Huang, J. Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: comparative evaluation of model-based linkage methods for affected sib pair data. *Hum. Hered.* **51**, 199–208 (2001).
14. Greenberg, D.A. Summary of analyses of problem 2 simulated data for GAW11. *Genet. Epidemiol.* **17** (Suppl. 1), S429–447 (1999).
15. Wang, K., Vieland, V. & Huang, J. A Bayesian approach to replication of linkage findings. *Genet. Epidemiol.* **17**, S749–754 (1999).
16. Gu, C. Province, M. & Rao, D.C. Meta-analysis of genetic linkage to quantitative trait loci with study-specific covariates: a mixed-effects model. *Genet. Epidemiol.* **17** (Suppl. 1), S599–604 (1999).
17. Oxman, A.D. & Guyatt, G.H. A consumer's guide to subgroup analyses. *Ann. Intern. Med.* **116**, 78–84 (1992).
18. Ioannidis, J.P. & Lau, J. Uncontrolled pearls, controlled evidence, meta-analysis, and the individual patient. *J. Clin. Epidemiol.* **51**, 709–711 (1998).
19. Yudkin, P.L. & Stratton, I.M. How to deal with regression to the mean in intervention studies. *Lancet* **347**, 241–243 (1996).
20. Easterbrook, P.J., Berlin, J.A., Gopalan, R. & Matthews, D.R. Publication bias in clinical research. *Lancet* **337**, 867–872 (1991).
21. Ioannidis, J.P. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *J. Am. Med. Assoc.* **279**, 281–286 (1998).
22. Cappelleri, J.C. *et al.* Large trials and meta-analysis of smaller trials: how do their results compare? *J. Am. Med. Assoc.* **276**, 1332–1338 (1996).
23. Hosmer, D.W. & Lemeshow, S. *Applied Logistic Regression* (John Wiley & Sons, New York, 1989).