Logical and methodological issues affecting genetic studies of humans reported in top neuroscience journals

Acronym: SQING (Study Quality in Neuro Genetics)

Clara R. Grabitz[a], Katherine S. Button[b], Marcus R. Munafò[c,d], Dianne F. Newbury[e], Cyril R. Pernet[f,] Paul A. Thompson[g], Dorothy V. M. Bishop[g*]

[a] Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands
[b]Department of Psychology, University of Bath, UK
[c]MRC Integrative Epidemiology Unit at the University of Bristol, UK
[d]UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, UK
[e]Department of Biological and Medical Sciences, Oxford Brookes University, UK
[f]Centre for Clinical Brain Sciences, Centre for Clinical Brain Sciences, Neuroimaging Sciences, The University of Edinburgh, UK
[g]Department of Experimental Psychology, University of Oxford, UK

*Corresponding author
Department of Experimental Psychology,
University of Oxford,
OX1 3UD, UK.
Tel: +44 01865 271369 Email: dorothy.bishop@psy.ox.ac.uk

Abstract

Genetics and neuroscience are two areas of science that pose particular methodological problems because they involve detecting weak signals (i.e., small effects) in noisy data. In recent years, increasing numbers of studies have attempted to bridge these disciplines by looking for genetic factors associated with individual differences in behaviour, cognition and brain structure or function. However, different methodological approaches to guarding against false positives have evolved in the two disciplines. To explore methodological issues affecting neurogenetic studies, we conducted an in-depth analysis of 30 consecutive articles in 12 top neuroscience journals that reported on genetic associations in non-clinical human samples. It was often difficult to estimate effect sizes in neuroimaging paradigms. Where effect sizes could be calculated, the studies reporting the largest effect sizes tended to have two features: (i) they had the smallest samples, and were generally underpowered to detect genetic effects; and (ii) they did not fully correct for multiple comparisons. Furthermore, only a minority of studies used statistical methods for multiple comparisons that took into account correlations between phenotypes or genotypes, and only nine studies included a replication sample, or explicitly set out to replicate a prior finding. Finally, presentation of methodological information was not standardized and was often distributed across Methods sections and Supplementary Material, making it challenging to assemble basic information from many studies. Space limits imposed by journals could mean that highly complex statistical methods were described in only a superficial fashion. In sum, methods which have become standard in the genetics literature – stringent statistical standards, use of large samples and replication of findings – are not always adopted when behavioural, cognitive or neuroimaging phenotypes are used, leading to an increased risk of false positive findings. Studies need to correct not just for the number of phenotypes collected, but also for number of genotypes examined, genetic models tested and subsamples investigated. The field would benefit from more widespread use of methods that take into account correlations between the factors corrected for, such as spectral decomposition, or permutation approaches. Replication should become standard practice; this, together with the need for larger sample sizes, will entail greater emphasis on collaboration between research groups. We conclude with some specific suggestions for standardized reporting in this area.

KEYWORDS: Genetics, Neuroscience, Methodology, Reproducibility, Power, Replication, Sample size, Multiple comparisons, Statistics, Neuroimaging, Effect size, Reporting guidelines, Sample selection

Introduction

Studies reporting associations in humans between common genetic variants and brain structure or function are burgeoning (Bigos, Hariri, & Weinberger, 2016). One reason is the desire to find ‘endophenotypes’ that provide an intermediate step between genetic variants and behaviour (Flint & Munafò, 2007); to this end, it is often assumed that brain-based measures will give stronger associations than observed behaviour because they are closer to the gene function. Furthermore, it is now cheaper and easier than ever before to genotype individuals, with many commercial laboratories offering this service, so neuroscientists interested in pursuing genetic studies need not have their own laboratory facilities to do this. The ease of undertaking genetic association studies is, however, offset by methodological problems that arise from the size and complexity of genetic data. As Poldrack et al (2017) cautioned with regard to neuroimaging data: "*the high dimensionality of fMRI data, the relatively low power of most fMRI studies and the great amount of flexibility in data analysis contribute to a potentially high degree of false-positive findings"*. When genetic approaches are combined with neuroscience methods, these problems are multiplied. Two issues are of particular concern.

The first issue is that the field of neuroscience is characterized by low statistical power (Button et al., 2013) where sample sizes are often too small to reliably detect effects of interest. Underpowered studies are likely to miss true effects, and where ‘significant’ effects are found they are more likely to be false positives. Where common variants are associated with behavioural phenotypes, effect sizes are typically very small; robust associations identified in genome-wide association studies (GWAS) typically account for less than 0.1% of phenotypic variance (Flint & Munafò, 2013). These reach genome-wide significance only when very large samples are used with this method. If we have a single nucleotide polymorphism (SNP) where a genetic variant accounts for .1% of variance (i.e., $r^2 = .001$), and we want to reliably detect an association of that magnitude, simple power calculations (Champely, 2016) show that we would need a total sample of 780 cases to detect the effect with 80% power at the .05 level of significance. If we had 200 participants (100 for each of two genotypes), then our power to detect this effect would be only 29%. Although it is often argued that effect sizes for neuroimaging phenotypes may be larger than for behavioural measures, a recent analysis by Poldrack et al (2017) suggests caution. They found that for a motor task that gives relatively large and reliable activation changes in the precentral gyrus, 75% of the voxels in that region showed a standardized effect size (Cohen's d) of less than one, and the median effect size was around .7; for other well-established cognitive tasks, the median effect sizes for a specified Region of Interest (ROI) ranged from .4 to .7. Furthermore, these effect sizes reflect within-subjects comparisons of the overall activation of task vs. baseline: when assessing differences in activation between groups, effect sizes can be expected to be smaller than this.

The second issue is that problems arise when there is a failure to appreciate that p-values are only interpretable in the context of a hypothesis-testing study (de Groot, 2014). Our knowledge is still limited, and many studies in this area are exploratory: insofar as there is a hypothesis, it is often quite general, namely that there may be a significant association between one of the genotypes examined and one or more phenotypes. Spurious findings are likely if there are many possible ways of analysing findings, and the measures or analyses are determined only after inspecting the data (Vul & Pashler, 2012). This leads to the twin problems of p-hacking (selecting and modifying analyses until a ‘significant’ p-value is found) and hypothesizing-after-results-are-known (Kerr, 1998), both of which render p-values meaningless. These practices are common but not easy to detect, although they may be suspected when there are numerous p-values just below a ‘significance’ threshold (Simonsohn, Simmons, & Nelson, 2015), or when the selection of measures or analyses has no obvious justification. One solution is to adopt a two-stage approach, where an association observed in an initial exploratory study (the "discovery" sample) is then tested in a more focused study that aims to replicate the salient findings in a fresh sample (the "replication" sample). This approach is now common in GWAS, after early association studies were found to produce numerous false positive findings. Before the advent of GWAS, the majority of reported associations did not replicate consistently (Sullivan, 2007). Most genetics journals now require that in order to be published, associations have to be replicated, and researchers have learned that large samples are needed to obtain adequate statistical power for replication (Lalouel & Rohrwasser, 2002) because initial reports overestimate true effect size, e.g. Behavioral Genetics (Hewitt, 2012). However, outside of GWAS, the importance of adequately powered replication is not always appreciated. As Poldrack et al (2017) noted, imaging genetics is '*a burgeoning field that has yet to embrace the standards commonly followed in the broader genetics literature*.'

An alternative approach to replication is to perform a statistical correction for the number of comparisons in an analysis. However, for this to be effective, the adjustment must be made for the multiplicity of potential analyses at several levels. Consider, for instance, a study where three SNPs are studied for association with measures of neural connectivity, based on four brain regions. If the SNPs are in linkage equilibrium (i.e., not associated) and the connectivity measures are uncorrelated, then it might seem that we could adequately control type 1 error by using a Bonferroni corrected p-value of .05/(3*4) = .004. However, suppose the researchers also study connectivity between brain regions, then there are six measures to consider (AB, AC, AD, BC, BD, CD). They may go on to test two models of genetic association (dominant and recessive) and further subdivide the sample by gender, increasing the number of potential comparisons to 3 * 6 * 2 * 2 = 72, and the Bonferroni-corrected p-value to .0007. Furthermore, we cannot compute this probability correctly unless all conducted tests are reported: if the authors remove reference to SNPs, genetic models, subgroups or phenotypes that did not yield significant results, then reported p-values will be misleading. In GWAS the finite search space (essentially the likely number of functional genetic variants in the human genome, estimated as around one million) means that a p-value threshold corrected for all possible tests can be calculated – in these studies, genome-wide significance for a single trait is typically set at 5×10^{-8} (Sham & Purcell, 2014).

Journal editors are becoming aware of problems of reproducibility in the field of neuroscience (Nicolas, Charbonnier, & Oliveira, 2015), many of which are reminiscent of earlier problems in the candidate gene era (Flint, Greenspan, & Kendler, 2010). The current study was designed to evaluate the extent to which these problems currently affect the field of human neurogenetics, and to identify instances of good practice that might suggest ways of overcoming the methodological and logical difficulties that researchers in this area face.

Study protocol

The protocol for this study was registered on Open Science Framework (https://osf.io/67jwb/). Many modifications were subsequently made in the course of collating studies for analysis, because papers or reported measures did not readily fit into the categories we had anticipated. Furthermore, the complexity of the methods used in many studies was such that it took substantial time to identify basic information such as effect sizes, which led to us focusing on a more restricted set of study features than we had originally planned. In addition, we added Cyril Pernet to the study team as it became clear that we needed additional expertise in neuroimaging methods to evaluate some of the papers. Departures from the protocol are noted below, with an explanation of each one.

Electronic search strategy

The search was conducted using the Web of Science database. We started with the 20 most highly-ranked journals in neuroscience and behaviour (source: https://www.timeshighereducation.com/news/top-20-journals-in-neuroscience-and-behaviour/412992.article). We then excluded journals that have a wide scope of subject matter (e.g., Nature, Proceedings of the National Academy of Sciences) and those that focus on review articles (e.g., Current Opinion in Neurobiology), which left 12 suitable journals to be used for the literature search. All of these publish articles in English only.

In our protocol, we planned to examine 50 publications, but we had underestimated the amount of time needed to extract information from papers, many of which were highly complex. When this became apparent, we decided that our resources would allow us to examine 30 publications in full, and so we restricted consideration to the most recent papers, starting with June 2016 and working backwards until 30 suitable papers were selected (initial search June 2016 – June 2011).

In order to identify relevant articles, the names of the 12 journals were coupled with topic-specific search terms. We limited the search to studies of humans, and used the following key terms:

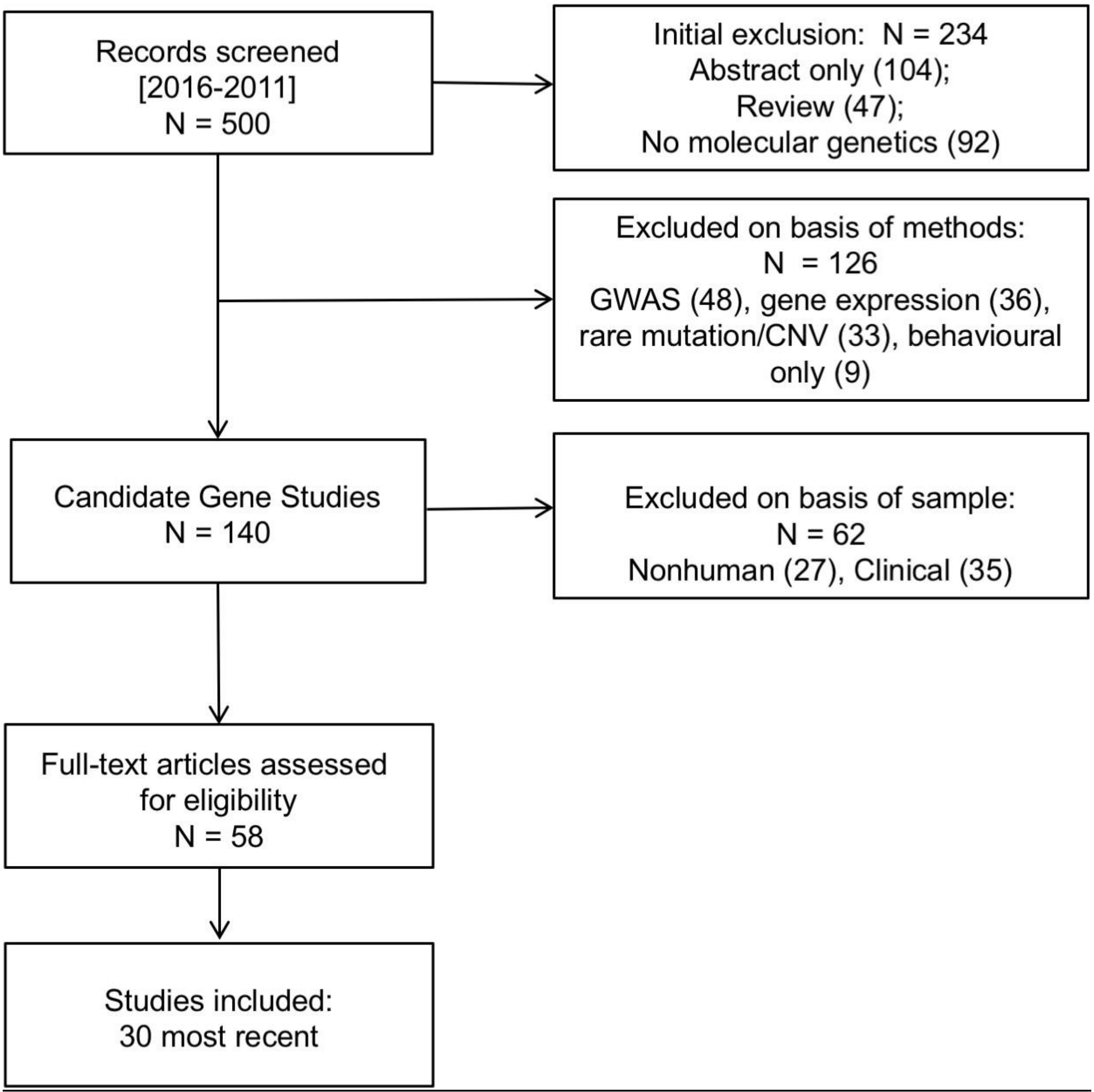
(Nature Neuroscience OR Neuron OR Annals of Neurology OR Brain OR Molecular Psychiatry OR Biological Psychiatry OR Journal of Neuroscience OR Neurology OR Journal of Cognitive Neuroscience OR Pain OR Cerebral Cortex OR Neurolmage) AND TOPIC: (genetic OR gene OR allele) AND TOPIC: (association) AND TOPIC: (cognition OR behaviour OR individual differences OR endophenotype) AND TOPIC: (human)

Selection criteria and data extraction

The first author screened abstracts found by the electronic search to identify relevant articles. The first and last author independently coded the first 500 articles and discussed sources of disagreement. This led to some refinement of the inclusion and exclusion criteria that had been specified in the original protocol, as described below (see Figure 1). The first 30 articles that met the final inclusion and exclusion criteria were fully reviewed and meta-data extracted (see below for details).

Figure 1

Flowchart showing stages of article selection


Records screened
[2016-2011]
N = 500
Initial exclusion: N = 234
Abstract only (104);
Review (47);
No molecular genetics (92)
Excluded on basis of methods:
N = 126
GWAS (48), gene expression (36), rare mutation/CNV (33), behavioural only (9)
Candidate Gene Studies
N = 140
Excluded on basis of sample:
N = 62
Nonhuman (27), Clinical (35)
Full-text articles assessed for eligibility
N = 58
Studies included:
30 most recent

<u>Inclusion criteria:</u>

- Candidate gene(s) study
- Studies predominantly focusing on healthy individuals. This includes population-based studies that may include individuals suffering from a disorder but where the phenotype of interest is a cognitive, behavioural or neuroimaging characteristic.

<u>Exclusion criteria:</u>

Original exclusion criteria specified in our protocol were:

- Review articles

- Genome wide association studies
- Studies predominantly focusing on genetic associations where the phenotype is a disease or disorder (e.g., neurodegenerative disease, neurodevelopmental disorder or psychiatric disorders)

Additional exclusionary criteria included after assembling pool of potential studies:

- Studies reporting an abstract only
- Studies solely on non-human species
- Studies solely focused on rare variants (i.e., those with a minor allele frequency less than 1%, or copy-number variants), because our focus was on common variation rather than disease, and rare variants and copy number variants require a different analytic approach.
- Studies focused solely on gene expression
- Studies with no molecular genetic content (e.g., twin studies)
- Analyses using polygenic risk scores

Data were extracted for the following characteristics:

1. Information about the study sample: the aim was to record information that made it possible to judge whether this was a clinical or general population sample, and if general population, whether a convenience sample or more representative
2. All SNPs that were tested
3. All measures of cognition, behaviour or neurological structure or function that were used as dependent variables
4. Sample size
5. Analysis method(s)
6. Any results given in terms of means and variance (SD or SE) on dependent measures in relation to genotype
7. Statistics that could be used to obtain a measure of association (odds ratios, regression coefficients, p-values, etc).

In our original protocol, we had planned also to evaluate the functionality of polymorphisms, to look for information on the reliability of phenotypes, and to evaluate the comprehensiveness of the literature review of each study, but the challenges we experienced in extracting and interpreting statistical aspects of the main results meant that we did not have sufficient resources to do this.

The information that we extracted was used to populate an Excel template for each study, which included information on sample size, corrections for multiple comparisons and whether or not a replication sample was included. The sample size was used to compute two indices of statistical power using the *pwr* package in R (R Core Team, 2016): (i) the effect size (r) detectable with 80% power; (ii) power of the study to detect an effect size (r) of .1.

We planned also to extract an effect size for each study, indicating the strength of genetic influence on the phenotype of interest. This proved difficult because many studies reported a complex range of results, with some including interaction effects as well as main effects of genotype. In addition, for studies reporting neuroimaging results, large amounts of data with spatial and temporal dependencies pose considerable challenges when estimating effect sizes, and so such studies were flagged as they often required alternative approaches.

To make the task of synthesizing evidence more tractable, we identified a ‘selected result’ for each study. To facilitate comparisons across studies and avoid the need for subjective judgement about the importance of different results, we identified this as the genotypic effect with the largest effect size (excluding any results from non-human species): this means that our estimates of study power give a 'best case scenario'. It also meant that in our summary template, study findings were often over-simplified, but we included a ‘comments’ field that allowed us to describe how this selected result fitted into the fuller context of the study. Our approach to computing a standard effect size is detailed below in the section on Analytic Approach.

In a further departure from our original protocol we sent the template for each study to the first and last authors with a request that they scrutinize it and correct any errors, with a reminder sent 2-3 weeks later to those who had not responded. Acknowledgement of the email was obtained from authors of 23 of 30 studies (77%), 19 of whom (63%) provided the requested information, either confirming the details in the template or making suggestions or corrections. The latter were taken into consideration in the summary of each study. We initially referred to the selected result with the largest genetic effect as a 'key result', and several authors were unhappy with this, as they felt that we should focus on the result of greatest interest, rather than largest effect size. We dealt with this by rewording and adding further explanation about other results in the study, noting when the selected result did not correspond to the author's main focus.

Simulations

We had not planned to include simulations in our protocol, but we found it helpful to write scripts to simulate data to explore two issues that arose. First, we considered how the false positive rate was affected when all three models: additive, dominant and recessive, were tested in the same dataset. Second, we considered how use of a selected sample (e.g., high ability students) might affect genetic associations when cognitive phenotypes were used.

Data extraction

Effect size: For each study, we aimed to extract an effect size, representing the largest reported effect of a genotype on a phenotype. For simple behavioural/cognitive phenotypes, it was usually possible to compute an effect size in terms of the correlation coefficient, *r*, which when squared provides the proportion of variance accounted for by genotype. The correlation coefficient is identical to the regression coefficient, *β*, when both predicted variable (y = phenotype of interest) and predictor (x = genotype) are standardized. For a standard additive genetic model with three genotypes (*aa, aA* and *AA*), the number of ‘risk’ alleles is the independent measure, so the regression tests for a linear increase in phenotypic score from *aa* to *aA* to *AA*. Where authors reported an unstandardized regression coefficient, *b*, the correlation coefficient, *r*, was obtained by the formula $r = b.s_x/s_y$, where s_x and s_y are the standard deviation for *x* (N risk alleles) and *y* (phenotype). Formulae from Borenstein et al (2009) were used to derive values of *r* when data were reported in terms Cohen’s *d*, odds ratios, or means and standard deviations by genotype. Where standard errors were reported, these were converted to standard deviations by the formula *SD = SE*sqrt(N).*

Two studies used structural equation modelling of relationships between variables, demonstrating that model fit was improved when genotype was incorporated in the model. In these cases, standardized parameter estimates or Pearson correlation coefficients relating genotype to phenotype were used to provide a direct measure of effect size (*r*).

For studies using phenotypic measures based on neuroimaging, an effect size can be estimated if an average measure of structure (e.g., grey or white matter volume) or function (e.g., or blood-oxygen-level dependent [BOLD] response) was taken from a brain region that was pre-defined in a way that was independent of the genetic contrast. For instance, the focus may be on a region that gave a significant group difference in a prior study, or the region may be chosen because it reliably gives strong activation on a task of interest. If several such regions are identified, then it is necessary to correct for multiple testing (see below), but the measure can be treated like any other phenotype when computing a standardized effect size, e.g. using the slope of the regression for three genotype groups in an additive model, to quantify how much variance in the neuroimaging measure is accounted for by genetic differences.

Few neuroimaging studies, however, adopted that approach. More commonly, they reported peak voxel statistics. This involves a statistical approach of searching for the voxel, or cluster of voxels, that gives the strongest effect, sometimes in a confined ROI, sometimes over many brain regions, and sometimes over the whole brain. The search volume can consist of tens or even hundreds of thousands of voxels. It is well-recognised in this field that correction of alpha levels needs to be made to control the rate of false positives, and a range of methods has been developed for this purpose. [1]

Although these methods make it possible to identify patterns of neural structure or function that differ reliably between groups, it is still not possible to derive a meaningful measure of effect size. This is because the focus is on just the subset of voxels that reached significance. As Reddan, Lindquist, & Wager (2017) put it, '*It is like a mediocre golfer who plays 5,000 holes over the course of his career but only reports his 10 best holes. Bias is introduced because the best performance, selected post hoc, is not representative of expected performance*.' In addition, the extent of the overestimate will depend on study-specific variables, such as the number of voxels considered and the size of clusters. Estimates of effects will also be distorted because of spurious dependencies in the data between true effects and noise (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). These problems are compounded by two further considerations. First, groups in genetic analyses are often unequal in size; where the dependent measure represents peak activation, the group with the biggest sample size and/or smaller variance in space will have a greater impact on the results. To continue the golfing analogy, if we compared two golfers on the basis of their best ten games, and one had played 100 games and the other only 20, then the one with the more games would look better, even if in fact there was no difference in skill.

It is not uncommon for researchers to use measures of peak activation but treat the resulting measures like more classic dependent variables (e.g., graphing means and standard errors for measures of activation across genetic groups, and reporting these along with corrected p-values). Such estimates are inaccurate, and possibly inflated, yet often these are the only kind of data available. Accordingly, where such approaches were adopted, we used the reported data to derive a ‘quasi effect size’, deriving *r* from means and SDs, but we treated these separately from other effect sizes, as they are likely to be distorted, and it is not possible to estimate by how much.

Analytic approach

[1] For more explanation, see Box 1 on Open Science Framework: osf.io/akuny

Our analysis was predominantly descriptive, and involved documenting the methodological characteristics of the 30 studies. In addition, we considered how effect size related to statistical power, and the methods used to correct for multiple comparisons.

Results

The genes and phenotypes that were the focus of each study are shown in Appendix 1 and full summary findings for each of the 30 studies are shown in Appendix 2. These are based on the templates that were sent to authors of the papers, but they have been modified on the basis of further scrutiny of the studies. In a preliminary check, we compared these papers to the set of 548 studies from the Neurosynth database that had been used by Poldrack et al (2017) to document trends in sample size for neuroimaging papers between 2011 and 2015. There was no overlap between the two sets.

Effect size of selected result in relation to sample size

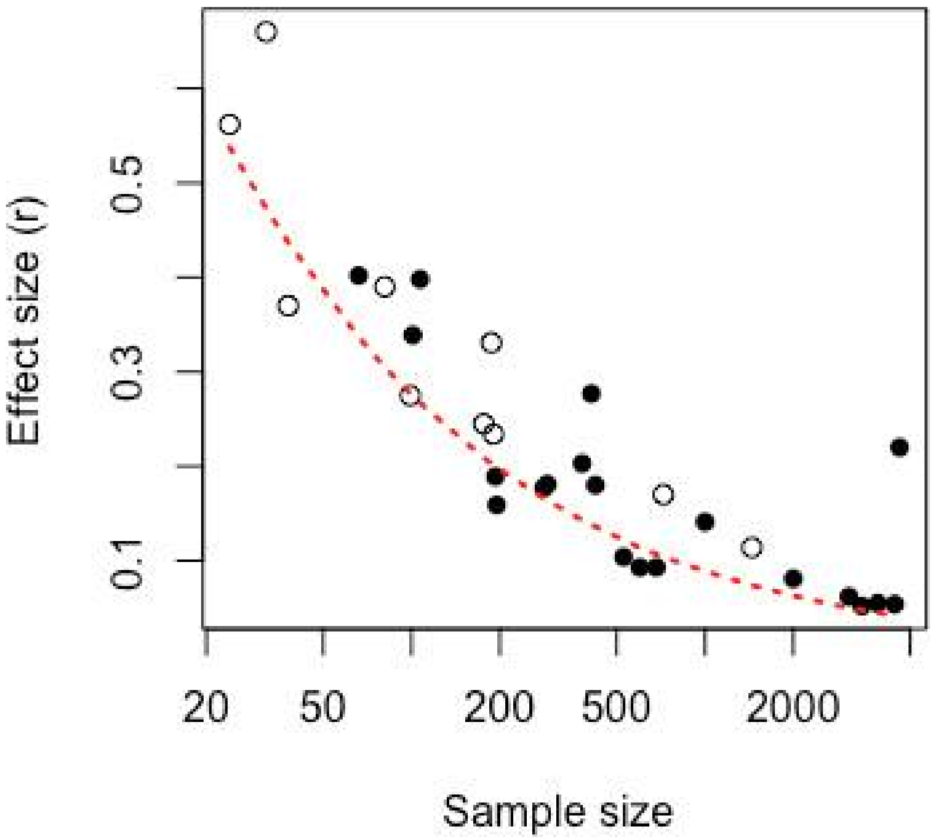
All the studies under consideration reported p-values, but only four explicitly reported conventional effect sizes (one as Cohen's *d*, and three as regression coefficients). Some fMRI studies mentioned 'effect size' or 'size of effect' when referring to brain activation, but this was on an arbitrary scale and therefore difficult to interpret. Nevertheless, we were able to compute an effect size from reported statistics for all studies that used behavioural (including cognitive) phenotypes, and quasi effect size (see above) for eight studies using neuroimaging phenotypes.

As noted, effect sizes of common genetic variants on behavioural or neurological phenotypes are typically small in magnitude. Where a research literature includes underpowered studies, effect size may be negatively correlated with sample size, reflecting the fact that small effects do not reach statistical significance in small samples and tend not to be published. This effect was apparent in the 30 papers included in our review. The relevant data are shown in Figure 2, where *r* is plotted against log sample size. Quasi effect sizes from neuroimaging studies are likely to be inflated, and so these are shown using different symbols.

Figure 2

Largest obtained effect size in relation to sample size (on log scale)

Quasi effect sizes (see text) shown as unfilled symbols. The red dotted line shows smallest effect size detectable with 80% power



The correlation between effect size and log sample size is -.85 (bootstrapped 95% CI = -.68 to .94) for the whole sample, and -.77 (bootstrapped 95% CI = -.38 to -.94) when ten neuroimaging studies with quasi effect sizes are excluded. It is clear from inspection that effect sizes (*r*) greater than .3 are seen only in studies where the total sample size is 300 or less. Only one study with a sample size of 500 or more obtained an effect size greater than .2. The largest reported effect size mostly clustered around the line corresponding to the effect detectable with 80% power: this makes sense insofar as studies are published only if they report statistically significant results. Thus, it is not that smaller studies show larger effects, but rather than in smaller studies, small effects would not be statistically significant, and so would tend to go unreported.

Corrections for multiple comparisons

The need to take multiple comparisons into account appears to be generally recognised: 23 of the 30 studies (77%) made some mention of this, although they varied widely in how they handled this issue. We had originally intended to simply report the number and nature of corrections used for multiple comparisons. However, this too proved complicated because there were many ways in which analytic flexibility could be manifest, with multiple comparison issues arising at several levels: in terms of analysis of subsamples, number of genetic models, number of polymorphisms, number of phenotypes, and, for neuroimaging studies, number of brain regions considered. In Table 1 we show for each study the numbers of comparisons at each level, as well as 'All Comparisons' which is the product of these. Matters are more complicated when there are dependencies between variables of a given type, as discussed in more detail below. Furthermore, it could sometimes be difficult to summarize the information, if certain phenotypes were assessed for just a subset of the sample, or were ambiguous as to whether they were phenotypes or moderators. In what follows, we first discuss multiplicity in terms of subgroups, then at genetic and phenotypic levels, before finally considering multiple comparisons in the context of neuroimaging studies.

(Table 1 in landscape format is at end of the paper)

Subgroups. In subgroup analysis, the association between genotype and phenotype is conducted separately for each subgroup (e.g., males and females). Typically, this is in addition to analysis of the whole sample with all subgroups included. Subgroup analysis is different from replication, where an association discovered in one sample is then confirmed in another, independent sample (see below). Most studies did not conduct any subgroup analysis, but four subdivided the participants by gender, one by ethnic group, one by age band, and one by psychiatric disorder in relatives.

It is well-known that deciding to analyse subgroups after looking at the data inflates type 1 error (Naggara, Raymond, Guilbert, & Altman, 2011), but there may be good *a priori* reasons for distinguishing subgroups. Subsampling by gender is justified where a relevant polymorphism is located on a sex chromosome, or where there are gender differences in the phenotype. Subsampling by ethnicity is generally advised to avoid spurious associations arising because of different proportions of polymorphisms in different ancestral groups (Tang et al., 2005) - known as population stratification. Nevertheless, subsamples will be smaller than combined samples, so power of the analysis is reduced, and furthermore each subsample included in an analysis will increase the likelihood of type 1 error unless the alpha level is controlled. Only two of the seven studies of subgroups made any adjustment for the number of subgroups.

Genetic variation. For the genotype part of genotype-phenotype association, there are two factors to take into account: (a) the number of polymorphisms considered; and (b) the number of genetic models tested

Number of polymorphisms: Polymorphisms are segments of DNA that take different forms in different people[2]. Most studies in our analysis investigated how phenotypes related to variation in one or more SNPs, with the number of SNPs ranging from one to 192.

Correlation between alleles at two or more genetic loci is referred to as linkage disequilibrium (LD). This can arise when loci are close together on a chromosome and so not separated by recombination events during meiosis, or it may be a consequence of population stratification, e.g., if certain genotypes are correlated with ethnicity or if there is assortative mating. Genetic variants that are inherited together on the same chromosome (i.e., from the same parent) give rise to combinations of alleles known as haplotypes. Rather than studying SNPs, some studies categorized participants according to haplotype status; this involves

[2] For more explanation, see Box 2 on Open Science Framework: osf.io/akuny

looking at the sequence of DNA in longer stretches of DNA, taking parent of origin into account.

Where polymorphisms are independent, a Bonferroni correction may be used by simply dividing the critical p-value by the number of SNPs (Clarke et al, 2011). For polymorphisms in linkage disequilibrium, the Bonferroni correction is overly-conservative. A range of methods has been developed to handle this situation and some of these are routinely output from genetic analysis software. For instance, the dimensionality of the data may be reduced by spectral decomposition, or by basing analysis on haplotype blocks rather than individual SNPs: these methods of data reduction are often incorporated as an additional step of correction for the effective number of comparisons, once the dimensionality had been reduced. Clarke et al (2011) noted that permutation methods are often regarded as the gold standard for correcting for multiple testing, but they are computationally intensive. Table 2 shows the different methods encountered in the 13 studies that reported analysis of more than one polymorphism. It is clear there is wide variation in the types of correction that are used, and some studies do not report any correction, despite studying two or more independent genotypes. Furthermore, correlations between polymorphisms were not always reported: in such cases, it was assumed they were uncorrelated.

Table 2

Correction for multiple testing in relation to genetic variants considered: 13 studies with 2 or more polymorphisms

	Correlated* Polymorphisms	Uncorrelated Polymorphisms
No	0	2
Bonferroni	2	2
Data Reduction**	3	0
Permutation	2	1

**Treated as correlated if authors reported greater than chance association between SNPs*
***e.g. using spectral decomposition to reduce dimensionality of data, or haplotype analysis*

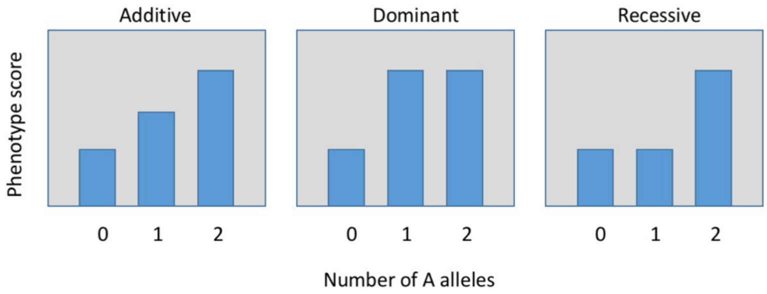
The majority of studies (N = 17) did not require any correction as only one SNP was reported. It is, of course, not possible to tell whether researchers tested a larger number of variants and selectively reported only those that reached statistical significance. A problem for the field is that it is difficult to detect this practice on the basis of published results. We know that dropping non-significant findings is a common practice in psychology (John, Loewenstein, & Prelec, 2012) and we may suspect selective reporting in studies where the choice of SNP seems arbitrary and unrelated to prior literature. We note below that requiring authors to report explicitly on whether all conducted tests were reported would ameliorate the situation. Furthermore, study pre-registration will remove uncertainty about which analyses were planned.

Eleven of the 13 studies that reported on two or more SNPs corrected for the number of genotypes tested, though two studies appeared to over-correct, by using a Bonferroni correction for correlated SNPs. The remaining studies used a range of approaches, some of which provided useful examples of how to deal effectively with the issue of multiple testing, as described further in the Discussion.

Genetic models: Consider a polymorphic SNP, with a major (more common) allele *A*, and a minor (less common) allele *a*, giving three possible genotypes, *AA, Aa* and *aa*. Let us suppose that *A* is the risk allele (i.e., associated with less optimal phenotype). There are three models of genetic association that are commonly tested: (i) additive model, tested by assessing the linear regression of phenotype on number of copies of allele *A*; (ii) a dominant effect, where *aa* differs from *AA* and *Aa*, with no difference between the latter two genotypes; and (iii) a recessive effect, where *AA* differs from *Aa* and *aa* (see Figure 3).

Figure 3

Schematic of three types of genetic model



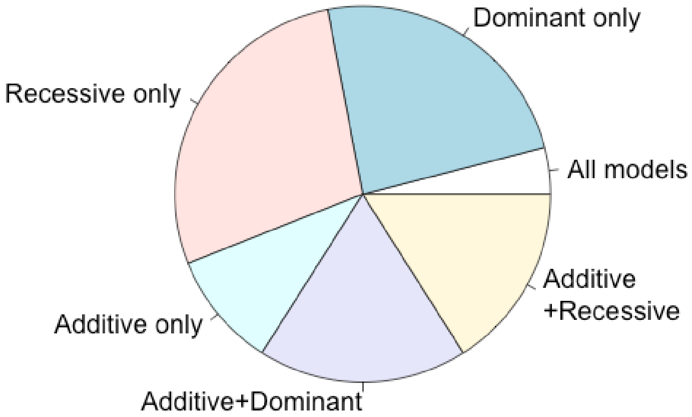
Some studies considered all three types of model, whereas others tested just one type of model. In other cases, the comparison was between 2 genotypes that corresponded to groups identified by length of tandem repeats, rather than base changes, and in one case a polymorphism on the X chromosome was considered in males, which gave a two-group comparison (base A vs base G) – because males have only one X chromosome.

There was only one study that explicitly tested three genetic models for each of several SNPs (additive, dominant, and recessive), and that study included a Bonferroni correction to adjust for this. This is, in fact, overly-conservative, as the predictions of an additive model partially overlap with those of recessive and dominant models. We devised a simulation to evaluate this situation. The phenotype was modelled as a random normal deviate, unrelated to simulated alleles at two loci for a SNP (A or a), so odds of obtaining a p-value < .05 for any one analysis should be one in 20. Regression analyses were run to look for an effect of number of A alleles (additive model), the effect of AA+Aa vs aa (dominant model), and the effect of AA vs Aa + aa (recessive model). Results indicated that adequate control for multiple comparisons is obtained by dividing the p-value by two (Figure 4). One study focused on interactions between two loci (epistasis) rather than main effects. Of the 28 remaining studies reporting just one genetic contrast per polymorphism, 17 reported results from additive genetic models (contrasting those with 0, 1 or 2 copies of an allele), nine reported only non-additive (dominant or recessive) models, and two included a mixture of additive and non-additive models, depending on the SNP. Of those reporting non-additive models, some justified the choice of model by reference to previous studies, but others grouped together heterozygotes and homozygotes with the minor allele for convenience because the latter group was too small to stand alone.

Figure 4

***Simulated data showing proportions of significant (p < .05) runs of a simulation that tests for all three genetic models when null hypothesis is true*.**

The total region of the pie corresponds to 10% of all runs (i.e., twice the expected 5%, but lower than the 14% that would be expected if the three models were independent). Note that we seldom see runs where both dominant and recessive models are significant, because they lead to opposite patterns of association (Figure 3), but it is not uncommon to see runs where both additive and recessive, or additive and dominant models are significant. For simulation code see: osf.io/4dymh.



Phenotypes. Phenotypes included measures of cognition, behaviour, psychiatric or brain functioning. For neuroimaging studies, the phenotypes included measures of brain structure or activation in response to a task. As described more fully below, the neuroimaging literature has developed particular strategies for dealing with the multiple contrasts issue; in Table 1, the number of brain regions is ignored when documenting the number of phenotypes. However, if brain activation was measured in several different tasks, then each task corresponded to a phenotype as defined for our purposes.

The simplest situation was where a phenotype was assessed using a behavioural or cognitive test that yielded a single measure, but this type of study was rare. Multiple phenotypic measures were common. As with genotypes, these were frequently correlated with one another, making Bonferroni correction too conservative, but studies often failed to report the extent of correlation between phenotypes. Often multiple measures were used to test the same construct, and so it is to be expected they would be intercorrelated: in such cases, if no mention is made of extent of intercorrelation, we record the correlation as 'unclear' in Tables 1 and 3. There was wide variation in the corrections used for the number of phenotypes. No correction was reported for 11 of 19 studies (58%) that included two or more phenotypes (see Table 3). In all cases, the phenotypes were correlated (or probably correlated): thus, conventional Bonferroni correction would have been too stringent.

Table 3

Correction for multiple testing in relation to whether behavioural phenotypes are correlated

	NA	Correlated	Probably correlated	Uncorrelated
None	0	2	9	0
Bonferroni	0	1	2	1
Permutation	0	1	0	2
Not needed	11	1*	0	0

**Mendelian randomization method*

Of the four studies using Bonferroni correction, three had correlated phenotypes, but one (study 9) took into account correlation between variables by reducing the denominator in the correction, though in what appeared to be an arbitrary fashion. More complex methods using permutation or bootstrapping were used in only three studies.

Neuroimaging phenotypes. In neurogenetics, the goal is to find structural or functional correlates of genotypes. It has long been recognised that neuroimaging poses multiple comparison problems of its own, since it typically involves gathering information from tens if not hundreds of thousands of voxels, corresponding to grey or white matter derived variables in the case of structural imaging (e.g., volume, thickness, anisotropy), or to proxies for underlying neural activity or connectivity in functional imaging. The spatial and temporal dependencies between voxels need to be taken into account.

The selection of a region of interest (ROI) is key. The commonest approach is to do a whole brain analysis. Some studies in our review selected specific regions and some assessed more than one region: in such cases, it is not sufficient to do statistical adjustments within the region – one still needs to treat each region as a new phenotype, with further correction applied to take the potential type I error inflation into account. The numbers for neuroimaging regions shown in Table 1 refer to the total ROIs that were considered in the analysis.

For the current analysis, we categorized neuroimaging papers according to whether they used a ROI specified a priori on the basis of previous research, with activation compared between genotype groups within that whole region. In such a case, it is possible to compare activation across genotypes to get a realistic effect size. However, as noted above, where the analysis involves first finding the voxel or cluster within a ROI that gives peak activation, and then comparing groups, it is not possible to accurately estimate effect sizes, because the method will capitalize on chance and so inflate these. Studies that adopted this approach are therefore flagged in Figure 1 as giving a 'quasi-effect size'.

Replication samples. We had originally intended to classify studies according to whether they included a replication sample, but this proved inadequate to handle the different approaches used in our collection of studies. As noted by Clarke et al (2011), a true replication uses the same SNPs and phenotypes as the original study, but in practice replication studies often depart from such fidelity and may study nearby variants of the same gene, or alternative measures of the phenotype. We categorized the replication status of each study as follows:

a) Study includes a full replication using comparable genotypes and phenotypes in the discovery and replication samples. This classification was less straightforward than it may appear. Consider, for instance, study 1: the replication sample included the same SNPs and measures from one of the same questionnaires as used in the discovery sample, but with a slightly different subset of items. In general, we treated a replication as full provided the measures were closely similar, so a case such as this would be regarded as a full replication.

b) Study includes a partial replication, but with some variation in genotypes or phenotypes in the discovery and replication samples.

c) Study is a direct replication of a previous study, so no replication sample is needed.

d) Study does not set out to replicate a prior study (though choice of phenotypes and genotypes is likely to be influenced by prior work) and does not include a replication sample

Even with this classification scheme, categorisation was not always straightforward. For instance, studies that did not include a replication sample would nevertheless usually aim to build on prior literature, and might replicate previous findings. These were categorised as 'prior' (option b) only if they were explicitly described as aiming to replicate the earlier work. We anticipated that replication samples would be more common in journals that regularly published genetics papers, where the need for replication is part of the research culture. Table 4 shows the number of papers according to replication status and journal. Note that there were three journals in our search for which no papers met our inclusion criteria in the time window we used: Nature Neuroscience, Neuroimage, and Brain.

Table 4

Number of studies including replication sample, by journal

	Yes	Partial	Prior*	No
Annals Neurol	0	0	0	1
Biol Psychiat	2	1	0	4
Cer Cortex	0	0	0	1
J Cog Neuro	0	1	0	2
J Neurosci	0	1	0	2
Mol Psychiat	4	1	2	4
Neurology	0	0	0	1
Neuron	0	0	0	1
Pain	1	0	0	1

**Study explicitly designed to replicate a prior finding*

Although the numbers are too small to be convincing, we may note that, in line with expectations, *Molecular Psychiatry*, which published the most studies in neurogenetics, was the journal with the highest proportion of studies including a replication, whereas neuroscience journals that did not have a genetics focus, and published few genetics studies, were more likely to publish studies without any replication sample.

Use of selected samples.

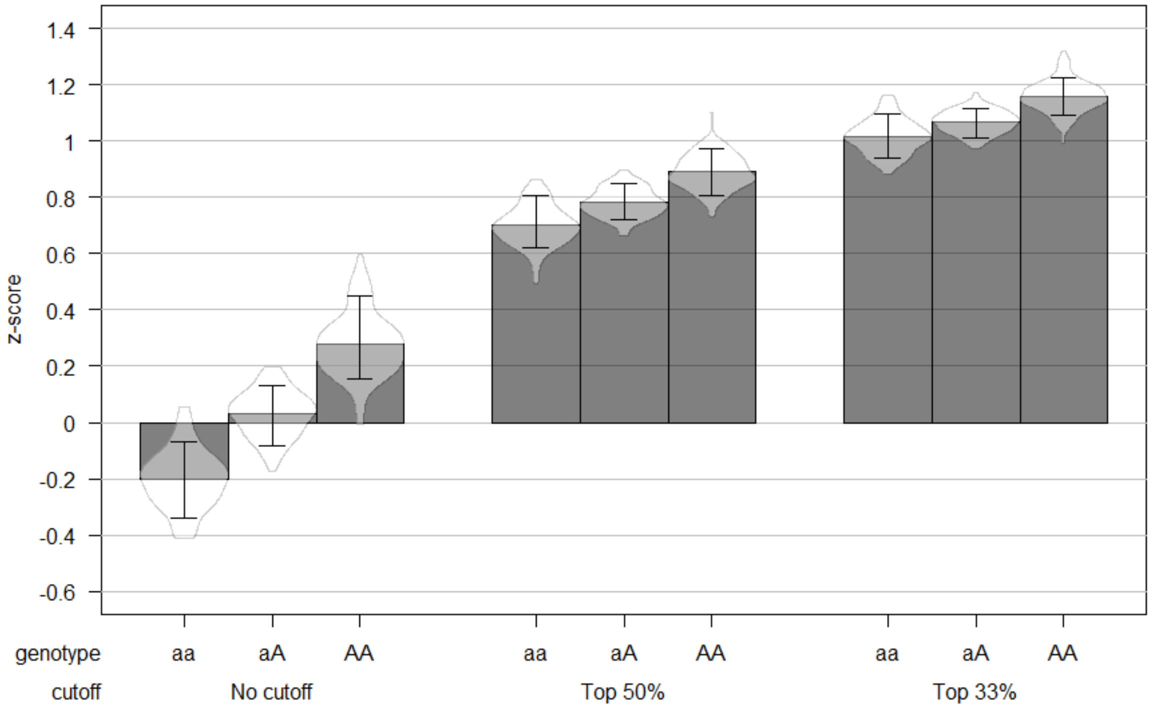
Some of the studies that we evaluated used samples from the general population, some used convenience samples, and some did not clarify how the sample had been recruited. Use of students has been criticised, on the grounds that people from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies are a very restricted demographic from which to make generalizations about human behaviour (Henrich, Heine, & Norenzayan, 2010). In the context of genetic research, however, other serious problems arise from the use of highly selected samples. Quite simply, if the phenotypic scores of a sample cover a restricted range, then power to detect genetic associations can be seriously affected.

We illustrate this with a simulation of an association between a genetic variant and a phenotype that has effect size of r = .2 in the general population. Let us assume that the minor allele frequency is .5, so the ratio of genotypes *aa, aA* and *AA* in the general population is 1:2:1. Now suppose we study a sample where everyone is above average on the phenotype, i.e. we only include those with positive z-scores. As shown in Figure 5, the effect of genotype on phenotype becomes substantially weaker. If we take an even more extreme group, i.e. the top third of the population, then the effect is no longer detectable in a moderate-sized sample. As also shown in Figure 5, as the association between genotype and phenotype decreases with selection, the ratio of the three genotypes changes, because those with the risk allele are less likely to be included in the sample. In fact, when there is strong selection, the effect of genotype will be undetectable, but the frequency of the three genotypes starts to depart significantly from expected values (see Figure 6).

Figure 5

Mean z-scores on a phenotype for three genotypes, when the true association between genotype and phenotype in the population is r = .2.

Data come from 10 000 runs of a simulation. The left hand panel shows the association in the full population; the middle panel shows means when the sample is taken only from those in the top 50% of the population on the phenotype measure, and the right-hand panel shows results when only the top third of the population is included. Ns are shown above the bars. As the selection becomes more extreme, the proportions of each genotype start to depart from the expected 1:2:1 ratio. The script 'simulating genopheno cutoffs.R' is available on: https://github.com/oscci/SQING_repo

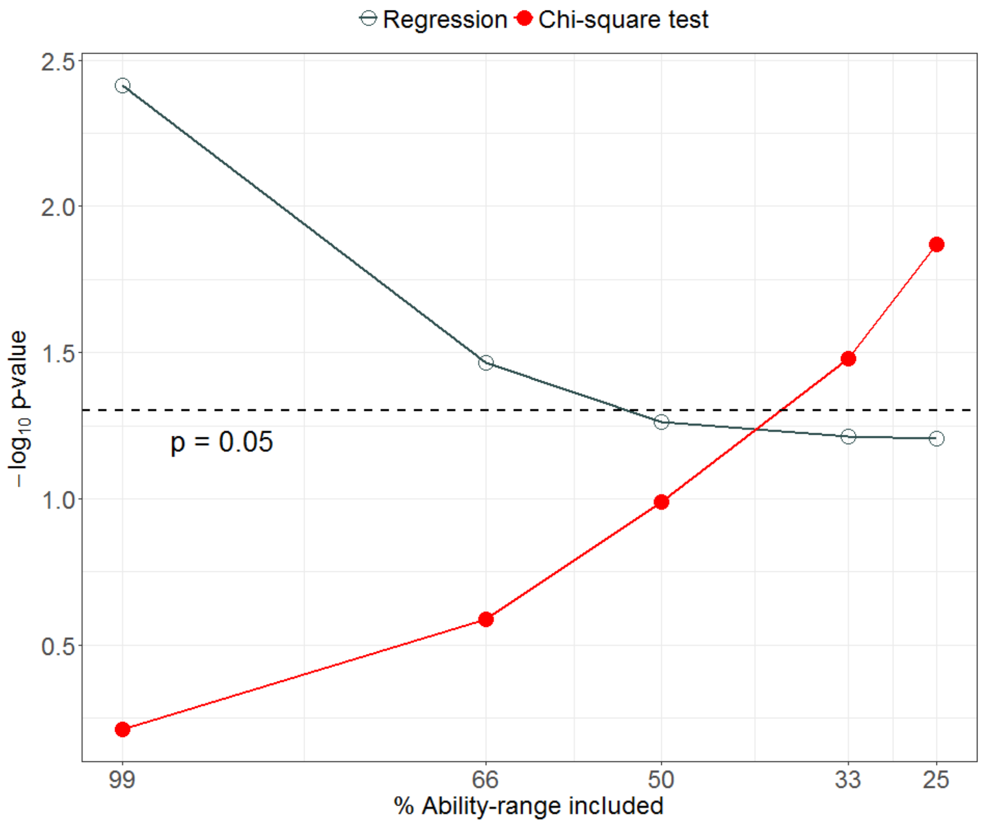


A corollary of this effect of sample selection is that moderate effect sizes on highly selected samples are implausible when the phenotype is related to the criterion for selection. This is because a moderate effect in a selected group would entail a much larger effect size in the general population, as well as skewing of the genotype distribution in the selected sample. Sample selection is therefore crucial. There may be situations when use of student samples is acceptable, because student status is unrelated to the phenotype of interest. However, where we are studying cognitive phenotypes, we stack the odds against finding associations with genotypes if we only study a high-functioning subset of the population. This can pose problems because, even when efforts are made to recruit a wide range of participants, those who volunteer tend to be biased towards more affluent and educated people (Rosenthal & Rosnow, 1975).

Figure 6.

Relationship between genotype and phenotype depending on how participants are selected.

The -log$_{10}$ p-values of the regression coefficient (blck unfilled circles) are shown for the association between genotype and phenotype for data simulated as in Figure 5, depending on whether the analysis is done on the whole population or a selected subset. The significance of the association decreases as the selection becomes stricter. The dotted line shows the log p-value corresponding to p = .05. When there is strong selection (inclusion only of top 33% or 25% of population on a phenotype z-score), there is significant departure from the expected 1:2:1 ratio of genotypes (as tested by chi-square test, red line).



Discussion

This in-depth analysis of 30 studies from top neuroscience journals complements other evaluations of data quality that have used text-mining methods to extract information from larger datasets. The studies varied widely in methods and phenotypes that were studied, with some providing good examples of best practice in the field. Nevertheless, we found that when neuroscience-oriented journals publish studies that include genetic analysis, they often fail to adopt the same stringent standards for sample size and replication as have become mandatory in the genetics literature.

An important limitation of our analysis is that we evaluated only 30 highly heterogeneous studies; it would not be realistic to assume that the proportion of studies with specific characteristics is representative of the field as a whole. Nevertheless, even with this small sample, it is clear that many genetic studies with neuro or behavioural phenotypes are underpowered and/or did not correct adequately for multiple testing, even though they were published in top journals.

Another limitation of our study is that it is based on just one 'selected result' per study, selected as the genetic association with the largest effect size. Many studies addressed questions that went beyond simple association between genotype and phenotype. Some considered the impact of functional groups of genes (e.g., Study 5), or looked at complex interactions between genetic variants, brain and behaviour phenotypes (e.g., Study 10). A few complemented studies of humans with animal models (e.g., Study 11). We note that studies that may look inconclusive when evaluated purely in terms of one selected result can compensate for this with converging evidence from a range of sources, and our analysis is not sensitive to this.

Despite this limitation of our approach, our analysis highlighted several issues that may need to be addressed in order for neurogenetic research to fulfil its promise.

Sample size and power. Sample sizes in this area are often too small to detect likely effects of genetic variation, particularly when neuroimaging phenotypes are used. A similar issue was highlighted for neuroimaging studies in general by Poldrack et al (2017), although they noted that sample sizes are now increasing as awareness of the limitations of small studies is growing. They concluded that sample sizes need to be justified by an a priori power analysis. The problem for researchers is that not only is power analysis complicated in neuroimaging (Mumford & Nichols, 2008), but also that these studies are difficult and time-consuming to conduct, and recruitment of suitable samples can take months if not years. However, Poldrack et al (2017) argued: '*We do not believe that the solution is to admit weakly powered studies simply on the basis that the researchers lacked the resources to use a larger sample*.' Instead, they recommend that, following the example of the field of genetics, researchers need to start working together in large consortia, so that adequate power can be achieved. A complementary approach is to pre-register a study, so that hypotheses, methods and analytic strategy are decided, and are publicly registered, before the data are collected; this can be invaluable in guarding against publication bias and the dangers of a flexible analytic pipeline. Some journals now offer Registered Reports, an approach where publication of a pre-registered study is offered, conditional on satisfactory reviews and adherence to the pre-registered protocol (Chambers, 2013).

An optimistic interpretation of the data in Figure 2 is that larger effect sizes are seen in smaller studies because these are studies that use highly specific measures of the phenotype that are not feasible with large samples. In particular, there is a widespread belief that neuroimaging will show stronger genetic effects than behavioural measures because it is closer to the mechanism of gene action. However, a more pessimistic interpretation is that where large effect sizes are seen in neuroimaging studies these are likely to be false discoveries arising from the use of small sample sizes with a very flexible analytic pipeline, and methods that tend to overestimate effect sizes.

Calculation of effect size. Our analysis highlighted another problem inherent in neuroimaging studies: the difficulty of specifying effect sizes. Lakens (2013) noted that effect size computations are not only crucial for establishing the magnitude of reported effects, but also for creating a literature that is useful for future researchers, by providing data in a format that can be combined with other studies in a meta-analysis, or which can be used to guide power calculations for future studies. Yet in neuroimaging, this is not standard. Indeed, only three of the 30 studies that we included explicitly mentioned effect sizes with a conventional interpretation of that term. This is consistent with a systematic review by Guo et al (2014) who found that only 8 of 100 neuroimaging studies reported effect sizes. When reported, effect sizes are typically shown for regions with the strongest effects and/or at the maximum voxel, leading to biased estimates. Correcting for multiple comparisons analyses further distorts these estimates, as the strongest voxels will be those with 'favourable' noise (i.e., spurious activity that adds to a true effect).

Correction for multiple comparisons. Most studies considered the issue of correction for multiple comparisons, but few fully corrected for all the tests conducted, taking into account the number of subgroups, genetic models, polymorphisms and phenotypes. Researchers appear to be aware of the multiple testing problem but there is not one good solution, and the impression was that sometimes authors thought they had done enough by applying standard corrections for fMRI, and did not need to correct for other aspects of the study. For instance, studies looking at correlations between genotypes or phenotypes in ROI, would have multiple comparisons procedures for whole brain analyses, but would either compute correlations for each ROI with no control, or conversely adopt a Bonferroni correction (which controls exactly the type one error rate) which is known to be is over-conservative .

In the field of genetics, a range of approaches has been developed for assessing associations when polymorphisms are not independent (i.e., in linkage disequilibrium); some of these, such as methods of data reduction by spectral decomposition or permutation tests, could be more widely applied (Clarke et al., 2011). For instance, extraction of latent factors from correlated phenotypes would provide a more sensitive approach for identifying genetic associations where a range of measures is used to index a particular construct, such as anxiety or memory.

Replication. Few studies included an independent replication sample, explicitly separating out the discovery and replication phases. This approach is now standard in GWAS, and has contributed to the improved reproducibility of findings in that literature. In principle this is a straightforward solution. In practice, however, it requires additional resources and means that studies take longer to complete. It also raises the possibility that findings in the discovery phase will not be replicated, in which case the overall results may be ambiguous. One solution to this problem is to apply a more stringent alpha level at the discovery phase than at the replication phase, and also to present results meta-analysed across both phases (Lander & Kruglyak, 1995). However, power calculations need to take into account the "winner's curse" phenomenon, which refers to the upward biasing of effect sizes when an original association emerged from a study considering many variants (Sham & Purcell, 2014).

Completeness of reporting. An unexpected feature of many of the studies that we analysed was the difficulty of finding the methodological information that we required from the published papers. Because there is no standard format for reporting methods, it could be difficult to know whether specific information (e.g., whether phenotypes were correlated) was simply omitted, or whether it might be found in Supplementary Material or figure legends, rather than the Methods section. Consequently, we had to read studies many times to find key information.

Most of the journals that we included had stringent length limits, or page charges, which might make it difficult for authors to report all key information. Exceptions were Neuroimage, Journal of Neuroscience, Pain and Neuron. It is noteworthy that in 2016 Neuron introduced new guidelines for Structured, Transparent, Accessible Reporting (STAR), and removed any length limit on Methods (http://www.cell.com/star-methods), with the goal of improving reproducibility of published studies.

Complexity of analyses. Several studies used complex analytic methods that were difficult to evaluate, despite the range of disciplinary expertise covered by the co-authors of our study. This in itself is potentially problematic for the field, because it means that reviewers will either decline to evaluate all or part of a study, or will have to take analyses on trust. One solution would be for journals to require researchers to make available all analysis scripts as well as raw data, so that others could work through the analysis.

Further considerations. We briefly mention here two additional issues that we were not able to evaluate systematically in the 30 papers that we considered, but are relevant for future research in this area.

i) Validity of genotype-phenotype association. We can be most confident that an association is meaningful if the genetic variant has been shown to be functional, with physiological effects that relate to the phenotype. Nevertheless, the ease of demonstrating functionality is much greater for some classes of variants than others. Furthermore, an association between a SNP and phenotype does not mean that we have found a functional polymorphism. Associated SNPs often lie outside genes and may be associated with phenotypes only because they are close to relevant functional variants – what has been referred to as 'indirect genotyping' (Clarke et al., 2011). Information about such variants can be valuable in providing landmarks to the key functional variant. With indirect genotyping, patterns of association may vary depending on samples, because different samples may have different patterns of linkage disequilibrium between genes and markers. It follows that a failure to replicate does not necessarily mean we have a false positive.

ii) Reliability and heritability of phenotypes. The phenotypes that are used in genetic association studies are increasingly diverse (Flint & Munafò, 2013). The idea behind the endophenotype concept is that a brain-based measure will be a more valid measure of the phenotypic effect of a genetic variant than other types of measure, because it is a more direct indicator of a biological effect. However, evidence for this assumption is lacking, and the strength of effects will depend on reliability as well as validity of phenotype measures. Quite simply, if a measure varies from one occasion of measurement to another, it is much harder to detect group differences even if they are real, because there will be noise masking the true effects. Therefore, it is advisable before embarking on a genetic association study to optimize – or at least assess - reliability of phenotypic measures. Psychometric tests typically are designed to take this into account and data on reliability will be available, but for most experimental and behavioural measures this is not the case. Furthermore, indices from functional imaging can vary from time to time (Nord, Gray, Charpentier, Robinson, & Roiser, 2017), and even structural imaging indices are far from perfectly reliable. Further problems occur when applying methods such as fMRI to the study of individual differences where people may differ in brain structure or trivial factors such as movement in the scanner, masking meaningful individual variation (Dubois & Adolphs, 2016).

As noted by Carter et al (2016) neurogenetic studies rely on the assumption that the phenotype is heritable. Yet, for many of the phenotypes studied in this field, evidence is lacking – usually because there are no twin studies using that specific phenotype. Heritability will be limited by reliability: a measure that shows substantial variation within the same person from one occasion to the next will not show good agreement between genetically related individuals.

Proposed reporting requirements for future articles

We conclude by making some suggestions that will make it easier for future researchers to understand neurogenetic studies and to combine these in meta-analyses, as detailed in Table 5. Ultimately, this field may need more formal reporting guidelines of the kind that have been designed to improve reproducibility of research in other areas, such as the guidelines for life sciences research introduced by Nature journals in 2015 (Nature Publishing Group, 2015), and the COBIDAS guidelines for MRI (Nichols et al., 2016). Making formal recommendations is beyond the scope of this article, but we suggest that if authors systematically reported this basic information in the Methods section of papers, it would be a major step forward.

Table 5: Key information for neurogenetic studies

Sample

- Provide a power calculation to determine the sample size. The usual recommendation is for 80% power based on estimated effect size, which may be based on results with this genetic variant in previous studies. If no prior effect size available, it is advisable to compute power with effect size no greater than $r = .2$, as few common genetic variants have effects larger than this. For neuroimaging studies, the application Neuropower (Durnez, Degryse, Seurinck, Moerkerke, & Nichols, 2015) is a user-friendly toolbox to help researchers determine the optimal sample size from a pilot study.
- Give total sample size. Where different numbers are involved at different points in a study, a flowchart is helpful in clarifying the numbers and reasons for exclusions.
- State how the sample was recruited, and whether they are representative of the general population for the phenotype of interest

Genetic variants

- State how many genetic variants were considered in the analysis
- List all genetic variants, regardless of whether they gave significant results
- Give background information indicating what is known about the genetic variants, what is known about the minor allele frequency, and whether they are functional.
- State whether or not the genetic variants are in linkage disequilibrium, and if so, how this is handled in the analysis
- State which genetic models were tested, and where genotypes are combined, whether this was to achieve a workable sample size, or whether the decision was based on prior research

Phenotypes

- State whether phenotypes are known to be heritable (e.g. using evidence from twin studies).
- Provide information on the test-retest reliability of the phenotype
- State whether phenotypes are inter-correlated
- Neuroimaging phenotypes involved many technical choices affecting the processing pipeline. Guidelines for reporting details of neuroimaging studies have been developed with the hope of improving reproducibility. The details of analytic information go beyond the scope of this paper, but useful information is given in Box 4 from Poldrack et al (Poldrack et al., 2017)

Analysis

- State which analyses were planned in advance. Post hoc analyses can be useful, but only if they are clearly distinguished from *a priori* hypothesis-testing analysis. Where there is a clear *a priori* hypothesis, consider pre-registering the study
- Describe the total number of independent tests that are conducted on the data. Describe the approach used to deal with multiple comparisons, bearing in mind that other approaches exist in cases where a Bonferroni correction is likely to be overconservative.
- Make scripts for processing the data openly available on a site such as Github or Open Science Framework. It is common for authors to describe complex methods that are hard even for experts to understand. By making scripts accessible, authors not only make their paper easier to evaluate, but they can also serve a useful training function, and facilitate replication

Results

- Do not rely solely on reporting derived statistics and p-values
- Show measures of central tendency and variation for each genotype group in relation to each phenotype, together with the effect size, where it is possible to compute this. Where the phenotype is categorical, report the proportions of people with each genotype who are in each category

References

Bigos, K. L., Hariri, A. R., & Weinberger, D. R. (2016). Neuroimaging genetics. In K. L. Bigos, A. R. Hariri, & D. R. Weinberger (Eds.), *Neuroimaging genetics: principles and practices* (pp. 1-14). New York: Oxford University Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*: John Wiley & Sons.

Brett, M., Penny, W. D., & Kiebel, S. J. (2003). Introduction to Random Field Theory. In R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan, K. J. Friston, C. J. Price, S. Zeki, J. Ashburner, & W. D. Penny (Eds.), *Human Brain Function, 2nd Edition*: Academic Press.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi:10.1038/nrn3475

Carter, C. S., Bearden, C. E., Bullmore, E. T., Geschwind, D. H., Glahn, D. C., Gur, R. E., . . . Weinberger, D. R. (2016). Enhancing the informativeness and replicability of imaging genomics studies. *Biological Psychiatry, e-print ahead of print*. doi:http://dx.doi.org/10.1016/j.biopsych.2016.08.019

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610. doi:10.1016/j.cortex.2012.12.016

Champely, S. (2016). pwr: Basic Functions for Power Analysis. R package version 1.2-0. Retrieved from https://cran.r-project.org/package=pwr

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols, 6*(2), 121-133. doi:http://www.nature.com/nprot/journal/v6/n2/abs/nprot.2010.182.html

de Groot, A. D. (2014). The meaning of “significance” for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica, 148*, 188-194. doi:http://dx.doi.org/10.1016/j.actpsy.2014.02.001

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences, 20*(6), 425-443. doi:10.1016/j.tics.2016.03.014

Durnez, J., Degryse, J., Seurinck, R., Moerkerke, B., & Nichols, T. E. (2015). Prospective power estimation for peak inference with the toolbox neuropower. *Front. Neuroinform. Conference Abstract: Second Belgian Neuroinformatics Congress*. doi:10.3389/conf.fninf.2015.19.00041

Flint, J., Greenspan, R. J., & Kendler, K. S. (2010). *How Genes Influence Behavior*: Oxford University press.

Flint, J., & Munafò, M. R. (2007). The endophenotype concept in psychiatric genetics. *Psychological Medicine, 37*, 163-180.

Guo, Q., Thabane, L., Hall, G., McKinnon, M., Goeree, R., & Pullenayegum, E. (2014). A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies. *Neuroimage, 86*, 172-181. doi:10.1016/j.neuroimage.2013.08.012

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. doi:https://doi.org/10.1017/S0140525X0999152X

Hewitt, J. K. (2012). Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits. *Behavior Genetics, 42*(1), 1-2. doi:10.1007/s10519-011-9504-z

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience, 12*(5), 535-540. doi:http://www.nature.com/neuro/journal/v12/n5/suppinfo/nn.2303_S1.html

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*. doi:10.3389/fpsyg.2013.00863

Lalouel, L. M., & Rohrwasser, A. (2002). Power and replication in case-control studies. *American Journal of Hypertension, 15*(2), 201-205.

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics, 11*, 241-247. doi:10.1038/ng1195-241

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage, 39*(1), 261-268. doi:10.1016/j.neuroimage.2007.07.061

Naggara, O., Raymond, J., Guilbert, F., & Altman, D. G. (2011). The problem of subgroup analyses: An example from a trial on ruptured intracranial aneurysms. *American Journal of Neuroradiology, 32*(4), 633-636. doi:10.3174/ajnr.A2442

Nature Publishing Group. (2015). Reporting Life Sciences Research. Retrieved from https://www.nature.com/authors/policies/reporting.pdf

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., . . . Yeo, B. T. T. (2016). Best practices in data analysis and sharing in neuroimaging using MRI, Annexe D. *bioRxiv, 054262*. doi:https://doi.org/10.1101/054262

Nicolas, G., Charbonnier, C., & Oliveira, J. R. M. (2015). Improving significance in association studies: a new perspective for association studies submitted to the Journal of Molecular Neuroscience. *Journal of Molecular Neuroscience, 56*(3), 529-530. doi:10.1007/s12031-015-0557-y

Nord, C. L., Gray, A., Charpentier, C. J., Robinson, O. J., & Roiser, J. P. (2017). Unreliability of putative fMRI biomarkers during emotional face processing. *Neuroimage*. doi:doi.org/10.1016/j.neuroimage.2017.05.024

Poldrack, R., Baker, C. I., Durnez, J., Gorgolewski, K., Matthews, P. M., Munafo, M., . . . Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*, 115-126. doi:http://dx.doi.org/10.1101/059188

Poline, J. B., & Mazoyer, B. M. (1993). Analysis of individual Positron Emission Tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *Journal of cerebral blood flow and metabolism, 13*(3), 425-437.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Reddan, M. C., Lindquist, M. A., & Wager, T. D. (2017). Effect size estimation in neuroimaging. *JAMA Psychiatry, 74*(3), 207-208. doi:10.1001/jamapsychiatry.2016.3356

Rosenthal, R., & Rosnow, R. (1975). *The volunteer subject*. New York: John Wiley.

Sham, P., & Purcell, S. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics, 15*, 335–346 doi:10.1038/nrg3706

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust To errors, fraud, and ambitious p-hacking, A reply To Ulrich and Miller (2015). *Journal of Experimental Psychology-General, 144*(6), 1146-1152. doi:10.1037/xge0000104

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage, 44*(1), 83-98. doi:10.1016/j.neuroimage.2008.03.061

Sullivan, P. F. (2007). Spurious genetic associations. *Biological Psychiatry, 61*, 1121-1126.

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L. R., Zhu, X. F., Brown, A., . . . Risch, N. J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *American Journal of Human genetics, 76*(2), 268-275. doi:10.1086/427888

Vul, E., & Pashler, H. (2012). Voodoo and circularity errors. *Neuroimage, 62*(2), 945-948. doi:10.1016/j.neuroimage.2012.01.027

Note: Scripts for the analyses reported in this paper are available on https://github.com/oscci/SQING_repo.

Acknowledgements

Dorothy Bishop and Paul Thompson are supported by Wellcome Trust Programme Grant no. 082498/Z/07/Z. Marcus R. Munafò is a member of the UK Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. Support from the Medical Research Council (MC_UU_12013/6) is also gratefully acknowledged.

Accompanying documents

Appendices 1 and 2 are available on Open Science Framework: osf.io/pex6w

Table 1

Corrections for multiple comparisons in relation to N subgroups, genetic models, polymorphisms, and imaging regions.

All combinations is the product of all of these. – denotes correlated variables; ~ denotes probably correlated

Study	Subgroups	Models	Polymorphisms	Phenotypes	Imaging regions	All Combinations	Correction method	Full correction
1	2	8	2-	4~	0	128	Bonferroni correction for 8 SNPs x 4 measures of phenotype x 2 genders.	Partial
2	1	3	1	5-	0	15	SEM with bootstrapping	Partial
3	1	2-	1	7~	0	14	None reported	No
4	2	1	1	7~	2	28	Imaging data FWE corrected, but no further correction reported for N overall analyses	No
5	2	50-	3-	1	0	300	Bonferroni separately for AA & EA ethnic groups; significance threshold for AA = 1.13 × 10-4 for 49 variants, 3 genetic models & 3 phenotypes; for EA=1.09 × 10-4 for 51 variants, 3 genetic models & 3 phenotypes)	Partial
6	1	1	1	1	2	2	Cluster-wise RFT for imaging data. No other corrections reported	No
7	1	9-	1	5-	0	45	Initial test of association of variants with categorical pain phenotype corrected using spectral decomposition	Partial

8	1	9	1	10-	0	90	P < .05 with no correction given strong prior evidence for all hypotheses	No
9	1	1	1	14-	0	14	P value of 0.01 was used instead of 0.05 to balance the risk of type I and type II errors,	Partial
10	1	1	1	19~	4	76	Separate Bonferronis: α level of .0055 for internal state analyses (9 time points); α level of .005 for perceptual ratings data (5 perceptual qualities for 2 types of stimuli)	Partial
11	1	1	1	1	3	3	None reported	No
12	1	36	1	1	0	36	36 SNPs captured the common haplotypic diversity of the TREM region: locus-wide Bonferroni-corrected p < 1.4 x 10-3; where genetic variant significantly associated with NP pathology, tested association with 5 secondary phenotypes, using Bonferroni-corrected p < 0.01	No
13	1	1	1	15~	0	15	None reported	No
14	3	1	1	1	3	9	Significance threshold was set to p .05, family-wise error corrected for multiple comparisons within our a priori defined anatomic regions of interest (FWEROI), the	Partial

							hippocampus and the pgACC.	
15	1	1	1	1	156	156	Sidak corrected significance level to maintain $\alpha = .05$ for testing 156 correlated outcomes (mean correlation $\rho = .25$) was determined at $p < 1.14 \times 10-3$	Yes
16	1	1	1	2-	0	2	Not needed; single polymorphism to test causal model using Mendelian randomisation	Yes
17	1	107-	1	2	0	214	Single step Monte Carlo permutation method	Yes
18	1	23-	1	4	4	368	Gene-wide significance was empirically determined with permutations that corrected for 23 SNPs (accounting for their LD structure), 4 ROIs, and the number of tests (main effects of SNPs, G×E interactions).	Yes
19	1	93-	1	1	0	93	No correction for multiple testing in initial sample because analysis conducted for discovery purposes.	No
20	1	1	1	14~	6	84	None reported	No
21	1	1	1	1	64	64	FDR-corrected p values for effect of APOE after adjusting for age, sex, and amyloid load (all ns)	Yes
22	1	10-	1	1	5	50	Significance level of 0.005 (0.05/10; Bonferroni corrected for the number of genetic tests	Partial

							conducted); No correction for number of ROIs	
23	2	1	1	7~	4	56	None reported	No
24	1	2	4-	1	1	8	None reported	No
25	1	1	1	1	1	1	Different for ROI and whole-brain; latter used fMRI significance measured at p < 0.05 family-wise error (FWE) corrected for multiple comparisons at the voxel level	Yes
26	1	2	1	3~	0	6	P < .05, with Bonferroni correction where appropriate. No Bonferroni for control analyses.	Yes
27	1	1	1	3~	0	3	Authors reply to query: "We did not correct for multiple testing as we only assayed 5-HTTLPR"	No
28	2	1	1	2~	4	16	Permutations with 100,000 iterations to control for hemisphere specific tests of VS BOLD response	Partial
29	1	1	1	10~	4	40	None reported	No
30	2	1	1	10	0	20	P-values adjusted for N inheritance modes. Considering the inter-correlation of 9 measures, reported nominal levels of significance. Bonferroni correction for 32 tests gave significance level of P=0.0016	Yes