Neuropsychopharmacology (2013) 38, 802–816

www.neuropsychopharmacology.org

Candidate Gene Studies of a Promising Intermediate Phenotype: Failure to Replicate

Amy B Hart[1], Harriet de Wit[2] and Abraham A Palmer*[,1,2]

[1]Department of Human Genetics, University of Chicago, Chicago, IL, USA; [2]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA

Many candidate gene studies use 'intermediate phenotypes' instead of disease diagnoses. It has been proposed that intermediate phenotypes have simpler genetic architectures such that individual alleles account for a larger percentage of trait variance. This implies that smaller samples can be used to identify genetic associations. Pharmacogenomic drug challenge studies may be an especially promising class of intermediate phenotype. We previously conducted a series of 12 candidate gene analyses of acute subjective and physiological responses to amphetamine in 99–162 healthy human volunteers (*ADORA2A*, *SLC6A3*, *BDNF*, *SLC6A4*, *CSNK1E*, *SLC6A2*, *DRD2*, *FAAH*, *COMT*, *OPRM1*). Here, we report our attempt to replicate these findings in over 200 additional participants ascertained using identical methodology. We were unable to replicate any of our previous findings. These results raise critical issues related to non-replication of candidate gene studies, such as power, sample size, multiple testing within and between studies, publication bias and the expectation that true allelic effect sizes are similar to those reported in genome-wide association studies. Many of these factors may have contributed to our failure to replicate our previous findings. Our results should instill caution in those considering similarly designed studies.
Neuropsychopharmacology (2013) **38,** 802–816; doi:10.1038/npp.2012.245; published online 16 January 2013

Keywords: *D*-amphetamine; intermediate phenotype; candidate gene; genetic association; replication

INTRODUCTION

A central goal of psychiatric genetics is to identify the small subset of polymorphisms that influence behavior out of the millions of polymorphisms that could, in principle, have such an effect. One approach is to focus on 'candidate genes,' which are typically genes for proteins involved in neurotransmission or with similarly well-understood functions. Many candidate gene studies have focused on intermediate phenotypes, for example, laboratory-based measures of normal behaviors. In contrast to endophenotypes, which must meet specific criteria (Gottesman and Gould, 2003), the term intermediate phenotype is sometimes used for traits that have not been formally shown to meet the criteria for endophenotypes (see Goldman and Ducci, 2007). It has been argued that intermediate phenotypes have a simpler genetic architecture than disease phenotypes, which would allow for the use of smaller samples (Goldman and Ducci, 2007). Drug response phenotypes, some of which can be considered intermediate phenotypes, have sometimes yielded large effect alleles (Daly, 2010), which stimulated our interest in intermediate phenotypes that focus on subjective drug responses.

*Correspondence: Dr AA Palmer, Department of Human Genetics, University of Chicago, 920 E 58th Street, CLSC-507D, Chicago, IL 60637, USA, Tel: +1 773 834 2897; Fax: +1 773 834 0505; E-mail: aap@uchicago.edu
Received 13 September 2012; revised 7 November 2012; accepted 26 November 2012; accepted article preview online 3 December 2012

Based on this reasoning, we investigated variability in acute response to a stimulant drug, *d*-amphetamine, in a large sample of healthy volunteers under highly controlled conditions. *D*-amphetamine response is known to be heritable in humans (Crabbe *et al*, 1983; Nurnberger *et al*, 1982) and behavioral responses to *d*-amphetamine are also heritable in mice (Alexander *et al*, 1996; Grisel *et al*, 1997; Kamens *et al*, 2005; Zombeck *et al*, 2010). Our study benefited from excellent experimental control and a reasonably large number of participants ($N=398$). The participants were normal-weight, psychiatrically and physically healthy young adults, with no history of substance dependence. We screened for drug use before each session, limited testing to the follicular phase in women, and counterbalanced the order of sessions. The study was double-blind, placebo controlled and included two (or in some participants, three) doses of the drug.

Over the 5 years that it took to collect these data, we conducted several interim analyses ($N=99$–162) that focused on carefully selected candidate genes: *ADORA2A* (Hohoff *et al*, 2005), *SLC6A3* (Hamidovic *et al*, 2010b; Lott *et al*, 2005), *BDNF* (Flanagin *et al*, 2006), *SLC6A4* (Lott *et al*, 2006), *CSNK1E* (Veenstra-VanderWeele *et al*, 2006), *SLC6A2* (Dlugos *et al*, 2007; Dlugos *et al*, 2009) *DRD2* (Hamidovic *et al*, 2009), *FAAH* (Dlugos *et al*, 2010), *COMT* (Hamidovic *et al*, 2010a), and *OPRM1* (Dlugos *et al*, 2011). These genes were examined using either the first 99 or the first 162 participants. The resulting publications have been cited over 200 times and have helped to inspire multiple similar

studies. In the present report, we have attempted to replicate our previously published associations in over 200 more recently collected participants that were recruited, screened and tested in an identical manner. Unlike many other attempts to replicate results from candidate gene studies, ours consists of multiple candidate genes, relatively large initial and replication cohorts and identical methodology. Thus, we avoided multiple sources of heterogeneity that are sometimes used to explain the failure of candidate gene studies to replicate.

MATERIALS AND METHODS

Here, we present the results of a new sample of young adults tested with three doses of *d*-amphetamine (0, 5, 10, 20 mg), under double-blind conditions exactly like those in the earlier studies. Because our goal was to replicate our previously reported associations, we first reanalyzed the data published previously for each gene (which we refer to as the 'original' sample), and then conducted an identical analysis with the new ('replication') sample using only the more recently collected participants. Replication was defined as obtaining a significant difference (in the same direction) in the replication sample when performing the same statistical test that was used in the original publication.

Study Design

Healthy young adults completed separate sessions during which they received placebo, 10 mg, or 20 mg of *d*-amphetamine. Some participants ($N = 299$) also participated in a fourth session with 5 mg. The study was performed under double-blind conditions with drug order counterbalanced. Earlier participants were genotyped at single SNPs and VNTRs, as well as using the Addictions Array (Hodgkinson *et al*, 2008). More recently, genotyping was performed for all 381 participants on the Affymetrix 6.0 array with imputation from the HapMap 3 (Frazer *et al*, 2007) and 1000 Genomes panels (Durbin *et al*, 2010), as previously described (Hart *et al*, 2012). VNTRs were directly genotyped in all 381 participants because they could not be reliably imputed using the SNP data.

Participants

The complete sample (ie, original and replication samples combined) consisted of 398 healthy volunteers aged 18–35 years old who were recruited locally and screened through a physical examination, electrocardiogram, modified Structured Clinical Interview for DSM-IV, psychiatric symptom checklist (SCL90) and health questionnaire that included sections on current and lifetime drug use. Exclusion criteria were: past year Axis I Disorder, history of mania or psychosis, less than a high-school level education, smoking >10 cigarettes per week, drinking more than three cups of coffee per day, lack of English fluency, a body mass index out of the range of 19–26 kg/m^2, any regular prescription medication except oral contraceptive or medical contraindication to amphetamine administration. Women not taking oral contraceptives were only tested in the follicular phase of their menstrual cycle (White *et al*, 2002). The final sample consisted of 381 participants (17 participants could not be included in the final analysis as discussed in Hart *et al*, 2012). Qualifying participants also provided a blood sample, or in some cases a saliva sample, for DNA analysis.

Phenotyping Procedure

Participants attended three or four 4-h sessions, conducted from 0900 to 1300 hours. They were tested individually in a comfortably furnished room located in the hospital. Sessions were separated by at least 48 h, and participants were instructed to abstain from drugs and alcohol for 24 h, nicotine for 12 h, and to fast for 12 h before each session. Before each session, participants provided urine (ToxCup, Branan Medical Corporation, Irvine, CA, USA) and breath samples (Alcosensor III, Intoximeters, St Louis, MO, USA; piCO+ Smokerlyzer, Bedfont, Rochester, UK) to confirm drug, alcohol, and nicotine abstinence, and female participants were tested for pregnancy. After compliance checks, participants completed subjective effects questionnaires (see below) and heart rate and blood pressure were recorded. They then ingested a capsule containing *d*-amphetamine (5, 10, or 20 mg) or placebo, under double blind conditions. During the next 3.5 h, participants relaxed in the laboratory, with reading materials or TV. They completed additional subjective effects measures 30, 60, 90, 150, and 180 min after the capsule, and physiological measures were also obtained at these times. At 120 min, they completed behavioral tasks described below. This study was approved by the Institutional Review Board of The University of Chicago and was carried out in accordance with the Helsinki Declaration of 1975.

Dependent Measures

Subjective measures consisted of three standardized questionnaires: the Profile of Mood States (POMS; Johanson and Uhlenhuth, 1980), Drug Effects Questionnaire (DEQ; Chait *et al*, 1985), and Addiction Research Center Inventory (ARCI; Martin *et al*, 1971). The POMS consists of 72 adjectives used to describe mood, ranging from 'not at all' (0) to 'extremely' (4). The subscales included from this questionnaire were 'Friendliness,' 'Elation,' 'Vigor,' 'Anger,' 'Anxiety,' 'Confusion,' 'Depression,' and 'Fatigue.' In some cases, the composite 'Positive Mood' (Elation − Depression) and Arousal [(Anxiety + Vigor)—(Fatigue + Confusion)] scales were analyzed. The DEQ consists of five 100 cm visual-analog scales describing five subjective responses to the drug: 'Feel Drug,' 'Want More,' 'Feel High,' 'Like Drug,' and 'Dislike Drug.' The ARCI is an empirically derived 52-item true/false questionnaire consisting of six subscales that measures effects of six classes of drugs: Amphetamine, Benzedrine, Marijuana, Lysergic Acid (LSD), Morphine-Benzedrine Group (MBG), and Pentobarbital-Chlorpromazine-Alcohol Group (PCAG). These measures were summarized in some analyses by either calculating the peak change score (PCS) or area under the curve (AUC). Behavioral tasks included the Stop Task (Logan *et al*, 1984), a measure of behavioral inhibition, and the Digit Symbol Substitution Task (DSST; Wechsler, 1958), a measure of motor-speed processing.

Genotyping and Quality Control

DNA was extracted from blood at the General Clinical Research Center at the University of Chicago. In the few cases where blood was not available, DNA was extracted from saliva samples with the Oragene OG-250 or OG-500 kit (Oragene, DNA Genotek, Kanata, Ontario, Canada). DNA from 15 participants could not be genotyped on the Affymetrix 6.0 array for technical reasons. We identified two participants who completed the study twice; we excluded their second sessions from the final data set. Thus, we had genotype and phenotype data from 381 participants in the final sample.

We were concerned that non-replication might reflect some systematic error in the replication sample (eg, misalignment of genotypes and phenotypes). Sample swaps can also often be detected as discordant genotypic and self-reported sex; however, we observed that genotypic sex was 100% consistent with self-reported sex.

Genotyping was performed in several stages throughout the course of the 5-year study. Participants were genotyped at single SNPs or VNTRs using PCR-based methods or on the Addictions Array (Hodgkinson *et al*, 2008); these genotypes were analyzed in our earlier publications. More recently, participants were genotyped on the Affymetrix 6.0 array as described in Hart *et al* (2012). We verified each individual's self-reported ancestry using the SmartPCA component of EIGENSOFT (Patterson *et al*, 2006), which generated ancestry principal components (PCs) that were included as covariates in reanalysis of original studies that included non-Caucasians.

We imputed non-genotyped SNPs with the IMPUTE2 software package (Howie *et al*, 2009), using the 1000 Genomes (Durbin *et al*, 2010) and HapMap3 (Frazer *et al*, 2007) phased genotypes as reference panels. Rs47958, rs6265, rs135745, rs36017, and rs4680 were genotyped on the Affymetrix 6.0 array, and rs5751876, rs1861647, rs4648317, rs12364283, rs3766246, rs2295633, and rs460000 were imputed. We checked the concordance of imputed genotypes by comparing them to the genotypes from the original studies. In all cases, the imputed genotypes had 96% or greater concordance with the direct genotypes, which demonstrated that these SNPs were well imputed.

VNTR Genotyping

SLC6A3 3′ UTR VNTR. Polymerase chain reactions were performed in a total volume of 25 μl containing: 1 × PCR buffer, 1.5 mM $MgCl_2$, 5% DMSO, 0.2 mM dNTPs, 0.4 mM of each primer (F: 5′-GGT GTA GGG AAC GGC CTG AGA-3′; R: 5′-CTT CCT GGA GGT CAC GGC TCA AGG-3′), 1.25 U *Taq* DNA polymerase (Fermentas, Glen Burnie, MD), and 100 ng DNA. Cycling conditions were 95 °C for 5 min, followed by 30 cycles of 94 °C for 30 s, 62 °C for 30 s, and 72 °C for 30 s. PCR products were resolved on a 2% agarose gel.

SLC6A4 Intron 2 VNTR. Polymerase chain reactions were performed in a total volume of 25 μl containing: 1 × PCR buffer, 1.0 mM $MgCl_2$, 0.2 mM dNTPs, 0.4 mM of each primer (F: 5′-TGG ATT TCC TTC TCT CAG TGA TTG G-3′; R: 5′-TCA TGT TCC TAG TCT TAC GCC AGT-3′), 1 U *Taq* DNA polymerase (Fermentas, Glen Burnie, MD), and 100 ng DNA. Cycling conditions were 95 °C for 2 min, followed by 35 cycles of 95 °C for 1 min, 62.5 °C for 1 min, and 72 °C for 2 min, with a final extension step of 72 °C for 10 min. PCR products were resolved on a 2% agarose gel.

Original Data Sets

In some cases the genotypes from prior studies were still available, which allowed us to exactly recreate the original analyses. In other cases, the original genotype information was no longer available and so we used genotypes obtained from the Affymetrix 6.0 array, by imputation, or by direct genotyping (VNTRs). In such cases, the sample was slightly different because DNA from a few of the earliest participants was no longer available. All phenotype data were available for reanalysis.

RESULTS

Table 1 summarizes the results of the original and the replication analyses. *The main conclusion is that none of our previous findings could be replicated using the newer data.* The demographic characteristics of the sample separated by 100's of sequentially tested participants are summarized in Table 2. This table shows that the sample was relatively uniform over the data collection period, except for race, which was mixed in the first 100 participants but was deliberately limited to Caucasian-only in the remainder; this issue is addressed in the section titled 'Population stratification analyses' (below). In the next sections, we summarize the findings of the original and the replication analyses for 10 genes that were the subject of 12 of our previous publications. For the purpose of this paper, we reanalyzed the original data using methods that were identical to the original publications, and then conducted the same analysis with the replication sample. The methods used in the original publications (including data reduction, selection of outcome measures, selection of covariates and data presentation) varied across studies, so these are described separately in each section. To facilitate comparison, we present the results in the same format that they appeared in the original publications. Results for the combined analyses (original and replication samples) are in Supplementary Table 1. Phenotypic means and standard deviations for each study are in Supplementary Table 2.

ADORA2A (Hohoff *et al*, 2005)

The original analysis for adenosine receptor genes (*ADORA1*, *ADORA2A*) consisted of 99 mixed-ancestry participants genotyped at three polymorphisms in *ADORA2A* (rs5760405, rs5751876, rs35320474) and one polymorphism in *ADORA1* (rs10920568). These genes were examined in relation to subjective (POMS subscales) and physiological responses to amphetamine using 3 × 5 × 3 repeated-measures ANCOVAs (Dose × Time × Genotype), with predrug scores used as covariates. *Post hoc* Dunnet's *t*-tests were used to assess the effect of specific genotypes. Hohoff *et al* (2005) identified a significant

Table 1 Summary of Recreated Original and Replication Analyses

Study	Gene	Polymorphism	Outcome measures	Recreated original *P* (dose × genotype ANOVA)	N (mixed/ Caucasian)	Replication *P* (dose × genotype ANOVA)	N (mixed/ Caucasian)
Hohoff *et al* (2005)[a]	ADORA2A	rs5751876	POMS Anxiety	0.041	98 (M)	0.624	281 (C)
Lott *et al* (2005)[a]	SLC6A3	3′ UTR VNTR	DEQ Feel	0.006	84 (M)	0.597[b]	273 (C)
			ARCI LSD	0.007	86 (M)	0.023[b]	272 (C)
			Diastolic BP	0.037	95 (M)	0.581[b]	284 (C)
Flanagin *et al* (2006)[a]	BDNF	rs6265 (pooled)	POMS Arousal	0.01	94 (M)	0.862	276 (C)
			ARCI BG	0.023	94 (M)	0.543	278 (C)
			Heart rate	0.023	94 (M)	0.791	290 (C)
Lott *et al* (2006)[a]	SLC6A4	Intron 2 VNTR	ARCI MBG	0.046	98 (M)	0.352	279 (C)
Veenstra-VanderWeele *et al* (2006)[a]	CSNK1E	rs135745	DEQ Feel	0.038	88 (M)	0.267	279 (C)
			ARCI MBG	0.008	89 (M)	0.109	278 (C)
Dlugos *et al* (2007)[a]	SLC6A2	rs47958	POMS positive mood	0.019[c]	90 (M)	0.562[c]	289 (C)
			POMS Elation	0.01[c]	90 (M)	0.278[c]	289 (C)
Dlugos *et al* (2009)	SLC6A2	rs36017	POMS Elation (gender cov)	0.154	156 (C)	0.57	170 (C)
			POMS vigor	0.041	156 (C)	0.334[b]	170 (C)
	SLC6A2	rs1861647	POMS Elation (gender cov)	0.137	155 (C)	0.875	170 (C)
			POMS vigor	0.006	155 (C)	0.116[b]	170 (C)
Hamidovic *et al* (2009)	DRD2	rs12364283 (pooled)	Stop RT	0.008	89 (C)	0.644[b]	122 (C)
Dlugos *et al* (2010)	FAAH	rs3766246	POMS Arousal (gender cov)	0.013	154 (C)	0.123[b]	173 (C)
			POMS fatigue	0.009[b]	155 (C)	0.187[b]	173 (C)
	FAAH	rs2295633	POMS Arousal (gender cov)	0.022	155 (C)	0.369[b]	173 (C)
			POMS fatigue	0.011	156 (C)	0.478[b]	173 (C)
Hamidovic *et al* (2010a)	COMT	rs4680	DSST	0.008	149 (C)	0.734	176 (C)
Hamidovic *et al* (2010b)	SLC6A3	rs460000 (pooled)	ARCI Amphetamine	0.015	152 (C)	0.03	169 (C)
			ARCI MBG	0.025	152 (C)	0.342[b]	169 (C)
Dlugos *et al* (2011)	OPRM1	rs510769	ARCI MBG	0.031[b]	162 (C)	0.204[b]	171 (C)
			ARCI A	0.019[b]	162 (C)	0.571[b]	171 (C)
	OPRM1	rs2281617 (pooled)	ARCI MBG	0.01[b]	162 (C)	0.340[b]	171 (C)
			ARCI BG (gender cov)	0.008[b]	162 (C)	0.923[b]	171 (C)

P-values for primary tests in the recreated and replication samples are shown here. *N* denotes sample size. 'Mixed' refers to samples that include both Caucasians and non-Caucasians.

[a]We were unable to locate the original genotype data files; therefore, we regenotyped subjects. However, because some DNA samples were unavailable, the sample size was smaller and therefore the values are slightly different from the initially published results.

[b]Greenhouse–Gessier correction when Mauchly's Test of Sphericity $P<0.05$.

[c]Kruskal–Wallis test.

Drug × Genotype interaction ($P=0.041$) between ratings on the POMS Anxiety subscale (PCS) and *ADORA2A* rs5751876 genotype. We obtained the same result (Figure 1a). *Post hoc* tests revealed that the rs5751876 T/T group had higher anxiety during the 10- and 20-mg sessions as compared with the C/T group (10 mg, $P=0.004$; 20 mg, $P=0.028$).

We conducted the same analysis used in the original publication in the replication sample ($N=281$). In the replication sample, the three genotype groups did not differ on the POMS Anxiety scale (Figure 1b), indicating a failure to replicate the original result from the original sample. Because we were concerned about the apparent differences in means between the original and replication samples, we plotted the distributions for each sample to verify that there was no overall difference in POMS Anxiety scores between the two samples. When the genotypic groups were combined, the distributions were very similar for the original sample and the replication sample (POMS Anxiety PCS, 20 mg; Supplementary Figure S1).

SLC6A3 (Lott *et al*, 2005)

The original analysis for the dopamine transporter gene (*SLC6A3*) consisted of 100 mixed-ancestry participants genotyped at the *SLC6A3* 3′ UTR VNTR polymorphism. Two common alleles exist (9-repeat and 10-repeat); four participants with rare alleles were excluded from the analysis. This polymorphism was examined in relation to subjective drug effects (POMS, DEQ, and ARCI subscales) and physiological responses to amphetamine using $3\times5\times3$ repeated-measures ANCOVAs (Dose × Time × Genotype), with predrug scores used as covariates. *Post hoc* *t*-tests were performed when a significant Drug × Genotype effect was found. Lott *et al* (2005) identified significant Drug × Genotype interactions between *SLC6A3* 3′ UTR VNTR genotype and ratings on the DEQ Feel ($P=0.006$) and ARCI LSD ($P=0.007$) subscales, as well as a significant association with diastolic blood pressure ($P=0.037$). We repeated this analysis and obtained the same results (Figure 1c; Supplementary Figures S2A and C).

Table 2 Demographic Characteristics of the Participant Sample Over Time

Demographic category	Demographic	Sample number (chronological)			
		1–100	101–200	201–300	301–398
General	Age (mean years ± SEM)	23.9 ± 0.41	22.7 ± 0.39	23.7 ± 0.32	22.8 ± 0.33
General	Gender (% male)	51	61	63	34
General	Education level—% high school or some college	54	46	26	46
General	Education level—% college degree	39	39	64	47
General	Education level—% advanced	7	15	10	7
General	BMI (mean + SEM)	22.6 ± 0.23	22.8 ± 0.22	22.4 ± 0.21	22.5 ± 0.21
Ancestry	% American Indian	1	0	0	0
Ancestry	% African American	20	0	0	0
Ancestry	% Asian	12	0	0	0
Ancestry	% Caucasian	54	95	97	95
Ancestry	% Hispanic	5	5	3	5
Ancestry	% More than one race	6	0	0	0
Ancestry	% Missing	2	0	0	0
Current drug use	Alcohol (mean drinks per week ± SEM)	4.22 ± 0.34	4.49 ± 0.38	5.89 ± 0.54	5.68 ± 0.47
Current drug use	Cigarettes (mean cigs per week ± SEM)	0.98 ± 0.24	0.53 ± 0.13	1.20 ± 0.28	1.05 ± 0.26
Current drug use	Caffeine (mean cups per week ± SEM)	9.19 ± 0.93	5.98 ± 0.53	8.31 ± 0.62	8.17 ± 0.55
Current drug use	Marijuana (mean times per month ± SEM)	0.97 ± 0.27	0.85 ± 0.18	1.72 ± 0.47	2.04 ± 0.53
Lifetime substance use (ever used recreationally)	Sedatives (% yes)	5	6	9	12
Lifetime substance use (ever used recreationally)	Stimulants (% yes)	17	25	25	38
Lifetime substance use (ever used recreationally)	Opiates (% yes)	10	16	20	27
Lifetime substance use (ever used recreationally)	Hallucinogens (% yes)	37	29	32	42
Lifetime substance use (ever used recreationally)	Inhalants (% yes)	12	10	8	13
Lifetime substance use (ever used recreationally)	Marijuana (% yes)	74	74	72	86

Demographic characteristics of the sample were computed over four phases of sample collection. Mean values are expressed as mean ± SEM.

Post hoc tests revealed that the 9/10 and 10/10 groups differed significantly for DEQ Feel at 20 mg compared with placebo (9/9: ns, 9/10: $P = 0.003$, 10/10: 4×10^{-9}); there was no significant difference for the 9/9 group. The same effect was seen for the ARCI LSD scale (9/9: ns, 9/10: $P = 2.9 \times 10^{-4}$, 10/10: $P = 1.13 \times 10^{-4}$). The 9/10 group differed significantly for diastolic blood pressure for the 20-mg session compared with placebo (9/9: ns, 9/10: $P = 0.004$, 10/10: ns). The 10/10 group did show significantly increased diastolic blood pressure at 10 mg compared with placebo ($P = 0.005$).

We conducted the same analysis used in the original publication in the replication sample ($N = 284$). In the replication sample, the all genotype groups showed significant responses to amphetamine on DEQ Feel (Figure 1d) and diastolic blood pressure (Supplementary Figure S2D), indicating a failure to replicate the results from the original sample. However, there was a modest but significant Dose × Genotype interaction for ARCI LSD ($P = 0.02$; Supplementary Figure S2B).

BDNF (Flanagin *et al*, 2006)

The original analysis for the brain-derived neurotrophic factor gene (*BDNF*) consisted of 99 mixed-ancestry participants genotyped at the val66met polymorphism (rs6265) in *BDNF*. Due to low minor allele frequency, the met/met group was pooled with the val/met heterozygote group in the original study. Associations of rs6265 with subjective mood scales and physiological measures were assessed with 3 × 5 × 2 repeated-measures ANCOVAs (Dose × Time × Genotype), with predrug scores used as covariates. Flanagin *et al* (2006) identified significant Drug × Genotype interactions between genotype at rs6265 and POMS Arousal ($P = 0.01$; Figure 1e) and ARCI BG ($P = 0.023$; Supplementary Figures S3A and C), as well as Dose × Genotype × Time interaction with heart rate (trend, $P = 0.057$; Supplementary Figures S3E and G), and we repeated this analysis and obtained the same results.

We conducted the same analysis used in the original publication in the replication sample ($N = 290$). In the replication sample, the two genotype groups did not differ on any measure (POMS Arousal: Figure 1f; ARCI BG: Supplementary Figures S3B and D; heart rate: Supplementary Figures S3F and H), indicating a failure to replicate the results from the original sample.

SLC6A4 (Lott *et al*, 2006)

The original analysis of the serotonin transporter gene consisted of 101 mixed-ancestry participants genotyped at the serotonin transporter (*SLC6A4*) Intron 2 VNTR and 5-HTTLPR polymorphisms. The two common Intron 2 VNTR alleles were analyzed (10 or 12 repeats) and individuals with rare alleles were excluded. These polymorphisms were analyzed in relation to subjective response to amphetamine

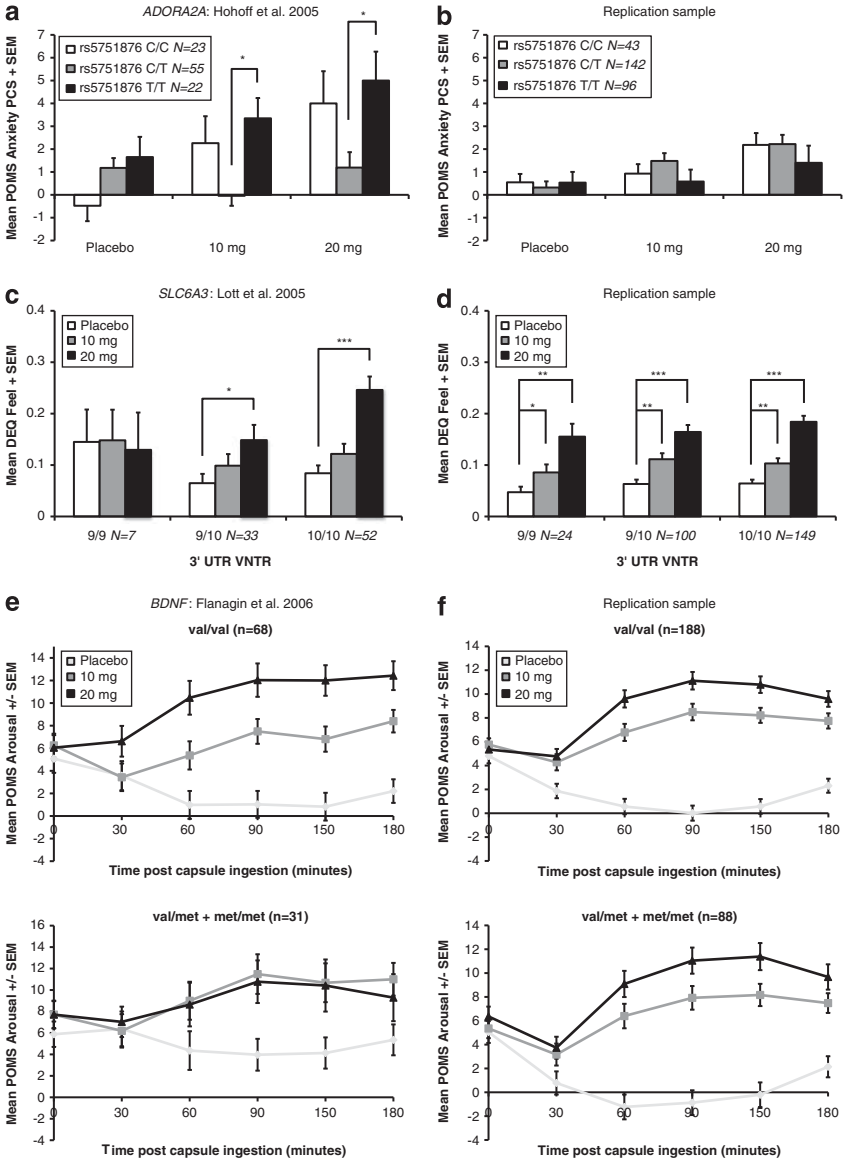

Figure 1 Replication results: *ADORA2A*, *SLC6A3*, and *BDNF*. (a) The recreated results from Hohoff *et al* (2005) (*ADORA2A*). T/T homozygotes had significantly higher ratings of Anxiety during the 10- and 20-mg drug sessions as compared with heterozygotes. (b) The results from the replication sample. No differences in anxiety between genotypes groups in the replication sample were observed. (c) The recreated results for DEQ Feel from Lott *et al* (2005) (*SLC6A3*). Participants with the 9/9 genotype showed less response to amphetamine when compared with heterozygotes and 10/10 homozygotes. (d) The results from the replication sample. All genotype groups showed the same pattern of response. (e) The recreated results from Flanagin *et al* (2006) (*BDNF*) for POMS Arousal. Val/val homozygotes showed more pronounced response to amphetamine when compared with val/met + met/met groups. (f) the results from the replication sample. Both the val/val and val/met + met/met groups showed similar responses to amphetamine. $^{*}P<0.05$, $^{**}P<0.001$, $^{***}P<10^{-4}$.

(DEQ Feel, POMS Anxiety, and ARCI MBG subscales) using $3\times5\times3$ repeated-measures ANCOVAs (Dose $\times$ Time $\times$ Genotype), with predrug scores used as covariates. *Post hoc* analyses were conducted with paired *t*-tests. Lott *et al* (2006) identified a significant Drug $\times$ Genotype interaction between ratings of ARCI MBG in response to 20 mg amphetamine. Our reanalysis of the original data produced the same result ($P=0.046$; Figure 2a). *Post hoc* tests with mean change scores from baseline revealed significantly greater mean ratings on the ARCI MBG subscale in the 10/10 group as compared with the 12/12 and 10/12 groups ($P=0.002$, $P=0.006$, respectively).

We conducted the same analysis used in the original publication in the replication sample ($N=279$). In the replication sample, we did not identify any difference between the three genotype groups (Figure 2b), indicating a failure to replicate the results from the original sample.

CSNK1E (Veenstra-VanderWeele *et al*, 2006)

The original analysis of the casein-kinase I epsilon gene (*CSNK1E*) consisted of 91 participants genotyped at three polymorphisms in *CSNK1E* (rs135745, rs1005473, rs199764). This polymorphism was analyzed in relation to subjective responses to amphetamine (DEQ Feel, POMS Anxiety, ARCI

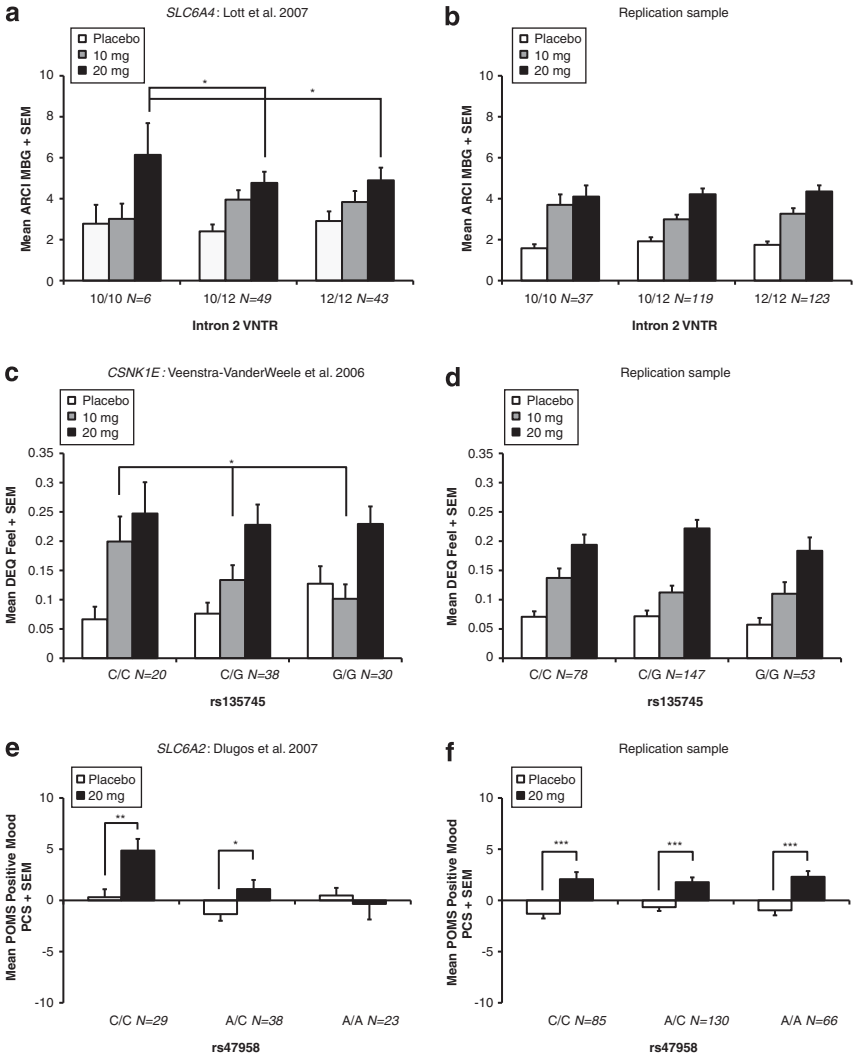

Figure 2 Replication results: *SLC6A4*, *CSNK1E*, and *SLC6A2*. (a) The recreated results from Lott *et al* (2006) (*SLC6A4*). 10/10 homozygotes showed greater subjective response to 20 mg amphetamine as compared with heterozygotes and 12/12 homozygotes. (b) The results from the replication sample. All genotype groups had similar levels of subjective response to 20 mg amphetamine. (c) The recreated results for the DEQ Feel subscale from Veenstra-VanderWeele *et al* (2006) (*CSNK1E*). The C/C group had greater ratings on the DEQ Feel subscale in response to 10 mg *d*-amphetamine as compared with C/G and G/G. (d) The results from the replication sample. No significant differences in subjective response to *d*-amphetamine were observed. (e) The recreated results from Dlugos *et al* (2007) (*SLC6A2*). C/C homozygotes and A/C heterozygotes had greater increases in subjective response to 20 mg amphetamine as compared with placebo; no such increase was seen in A/A homozygotes. (f) The results from the replication sample. All genotype groups showed similar increases in response to drug. $^{*}P<0.05$, $^{**}P<0.001$, $^{***}P<10^{-4}$.

MBG subscales) using $3\times5\times3$ repeated-measures ANOVAs (Dose $\times$ Time $\times$ Genotype). Predrug scores were subtracted from the score at each time point to yield change scores. *Post hoc* analyses were conducted to assess the effect of dose on the Genotype $\times$ Dose interaction; these consisted of $2\times5\times3$ repeated-measures ANOVAs (Dose $\times$ Time $\times$ Genotype). Veenstra-VanderWeele *et al* (2006) identified significant Drug $\times$ Genotype interactions between DEQ Feel and ARCI MBG change scores and genotype at rs135745 ($P=0.038$; $P=0.008$), an effect that was specific to the 10-mg dose ($P=0.001$; $P=0.004$); we repeated this analysis and obtained the same result (Figure 2c; Supplementary Figure S4A).

We conducted the same analysis used in the original publication in the replication sample ($N=279$), but did not identify any differences between the three genotype groups on any measure (Figure 2d; Supplementary Figure S4B), indicating a failure to replicate the results from the original sample.

SLC6A2 (Dlugos *et al*, 2007)

The original analysis of the norepinephrine transporter gene (*SLC6A2*) consisted of 99 participants genotyped at eight SNPs in *SLC6A2* (rs35915, rs168924, rs168924, rs2242446, rs36017, rs2270935, rs47958, rs171798). These SNPs, along with eight haplotypes comprised of these SNPs, were examined in relation to subjective responses to amphetamine using the non-parametric Kruskal–Wallis test. Dlugos *et al* (2007) identified a significant association between ratings of POMS Positive Mood (PCS; $P=0.019$) and POMS Elation (PCS; $P=0.01$) following amphetamine

administration (20 mg) and genotype at rs49758, and we repeated this analysis and obtained the same results (Figure 2e). *Post hoc* tests revealed significantly higher ratings of Positive Mood in response to 20 mg amphetamine in the C/C group ($P = 0.003$) and the A/C group ($P = 0.007$), but not in the A/A group ($P = 0.6$), as well as significantly higher ratings of Elation in the C/C ($P = 1.34 \times 10^{-4}$) and A/C ($P = 0.001$), but not the A/A group ($P = 0.4$). Additionally, Dlugos *et al* found that the rs36017–rs2270935–rs47958 GCC and CCA haplotypes were significantly associated with ratings of POMS Positive Mood (20 mg PCS), and we repeated the analysis and obtained the same results (GCC, $P = 0.032$; CCA, $P = 0.016$).

We conducted the same analysis used in the original publication in the replication sample ($N = 289$), but in the replication sample, the genotype groups did not differ on any measure (Figure 2f). Furthermore, neither the GCC haplotype ($P = 0.453$) nor the CCA haplotype ($P = 0.573$) was associated with ratings of POMS Positive Mood in the replication sample. Thus, we failed to replicate the results from the original sample.

SLC6A2 (Dlugos *et al*, 2009)

The original analysis of the norepinephrine transporter gene (*SLC6A2*) consisted of 162 Caucasian participants genotyped at 11 SNPs in *SLC6A2* (rs2397771, rs3785143, rs192303, rs36024, rs36021, rs3785152, rs36017, rs10521329, rs3785155, rs1861647, rs5569). These SNPs, along with two haplotypes comprised of these SNPs, were analyzed in relation to POMS Elation and Vigor subscales (PCS) in response to amphetamine using 3×3 repeated-measures ANOVAs or ANCOVAs (Dose $\times$ Genotype). Gender was used as a covariate in the analyses of POMS Elation, as it was seen to be associated with this subscale. *Post hoc* one-way ANOVAs were performed. Dlugos *et al* (2009) identified associations between POMS Vigor and Elation following amphetamine administration and *SLC6A2* SNP genotypes. We obtained the same results using the same data (Figure 3a; Supplementary Figures S5A, C and E). Specifically, significant Drug $\times$ Genotype interactions were identified for rs36017 and Vigor ($P = 0.041$) and rs1861647 and Vigor ($P = 0.006$). Although not statistically significant, trends were seen for rs36017 and Elation ($P = 0.154$) and rs1861647 ($P = 0.137$). *Post hoc* analyses revealed that individuals with the C/C genotype at rs36017 had significantly higher ratings of POMS Vigor following 20 mg amphetamine when compared with the C/G and G/G groups ($P = 0.003$, $P = 0.019$, respectively; Figure 3a). Similarly, this group had significantly higher ratings of POMS Elation in response to 20 mg amphetamine when compared with the C/G group ($P = 0.013$; Supplementary Figure S5A). The rs181647 A/A group had significantly higher ratings of POMS Vigor ($P = 0.01$; Supplementary Figure S5C) and POMS Elation ($P = 0.017$; Supplementary Figure S5E) when compared with the G/G group. Additionally, Dlugos *et al* found that the rs36017–rs10521329–rs3785155 CCG and rs1861647–rs5569 GC haplotypes were significantly associated with ratings of POMS Vigor (20 mg PCS), and we repeated this analysis and obtained similar results (rs36017–rs10521329–rs3785155 CCG, $P = 0.097$; rs1861647–rs5569 GC, $P = 0.0142$).

We conducted the same analysis used in the original publication in the replication sample ($N = 170$), but did not identify any differences between the three genotype groups for either SNP on any measure (Figure 3f; Supplementary Figures S5B, D and F). Neither the CCG haplotype ($P = 0.667$) nor the GC haplotype ($P = 0.571$) was associated with ratings of POMS Vigor in the replication sample. Thus, we failed to replicate the results from the original sample.

DRD2 (Hamidovic *et al*, 2009)

The original analysis of the dopamine D2 receptor gene (*DRD2*) consisted of 93 Caucasian participants genotyped at 12 SNPs in *DRD2* (rs2242592, rs1079596, rs1125394, rs27471857, rs4648317, rs4350392, rs1799978, rs12364283, rs71003679, rs4648318, rs4274224, rs4581480). In addition to 10 and 20 mg of amphetamine, this study also included the 5-mg dose. These SNPs were analyzed in relation to performance on the Stop Task following amphetamine administration using 4×3 repeated-measures ANOVAs (Dose $\times$ Genotype). Paired-samples *t*-tests were used to assess the effect of drug on each genotype group when a significant Drug $\times$ Genotype interaction was found. Hamidovic *et al* (2009) identified a significant Drug $\times$ Genotype interaction between genotype at rs12364283 and scores on the Stop Task in response to amphetamine ($P = 0.008$), and we repeated this analysis and obtained the same result (Figure 3c). *Post hoc* tests revealed that amphetamine decreased stop reaction time (Stop RT) in the A/A group as compared with placebo (5 mg, $P = 0.02$; 10 mg, $P = 0.001$; 20 mg, $P = 0.05$), but did not decrease Stop RT in the combined A/G + G/G group, and the 10-mg amphetamine dose significantly increased Stop reaction time compared with placebo in the combined A/G + G/G group ($P = 0.043$; Figure 3c).

We conducted the same analysis used in the original publication in the replication sample ($N = 122$), which was reduced in size because we excluded participants that possessed low quality Stop RT data. We did not identify any differences between the genotype groups (Figure 3d), indicating a failure to replicate the results from the original sample.

FAAH (Dlugos *et al*, 2010)

The original analysis of the fatty acid amide hydrolase gene (*FAAH*) consisted of 159 Caucasian participants genotyped at four SNPs in *FAAH* (rs6703669, rs3766246, rs324420, rs2295633). These SNPs were analyzed in relation to subjective responses to amphetamine (POMS Arousal, Fatigue subscales; AUC) using 3×3 repeated-measures ANOVAs/ANCOVAs (Dose $\times$ Genotype). *Post hoc* analyses were carried out with one-way ANOVAs. Gender was used as a covariate in the analyses of POMS Arousal, as it was found to be associated with this subscale. Dlugos *et al* (2010) identified associations between two SNPs in *FAAH* and scores on the POMS Arousal and Fatigue subscales in response to amphetamine. Significant Drug $\times$ Genotype interactions were found for rs2295633 and POMS Arousal ($P = 0.02$) as well as Fatigue ($P = 0.01$). We repeated this analysis and obtained the same results (Figure 3e;

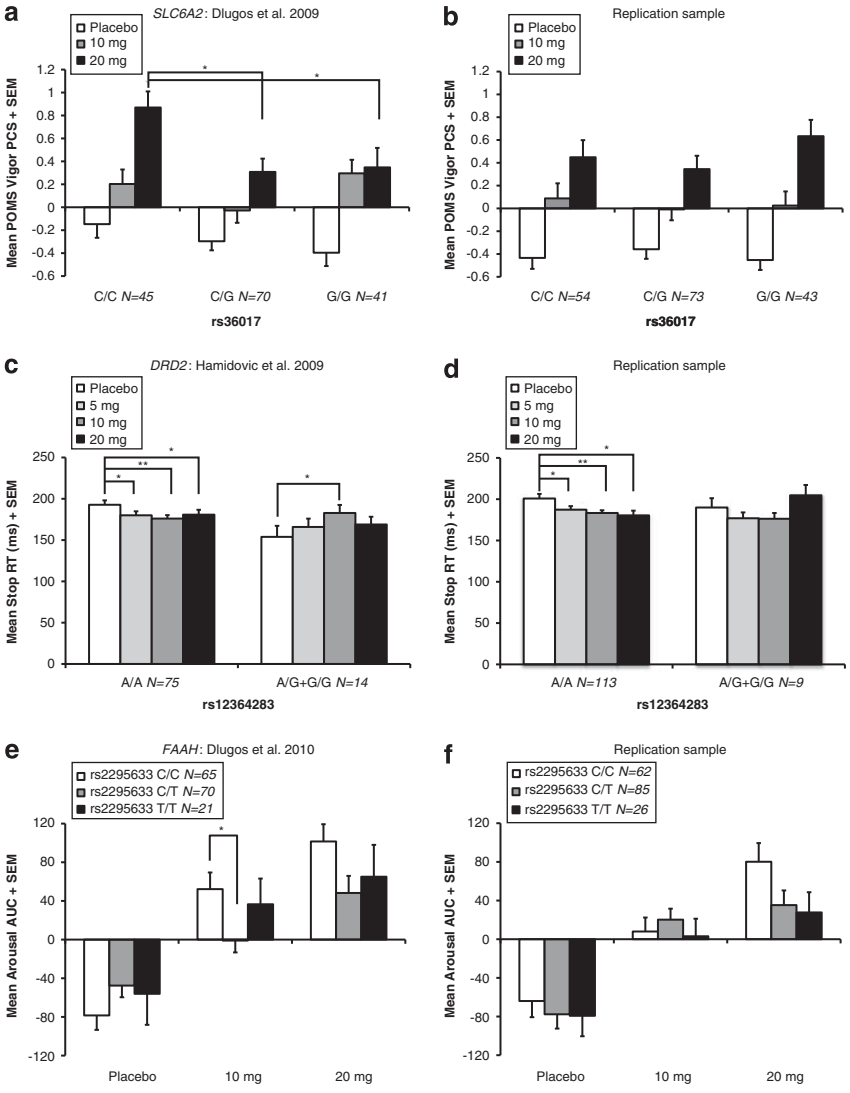

Figure 3 Replication results: *SLC6A2*, *DRD2*, and *FAAH*. (a) The recreated results from Dlugos *et al* (2009) for rs36017 (*SLC6A2*). C/C homozygotes had increased ratings of Vigor in response to 20 mg amphetamine in comparison to heterozygotes and G/G homozygotes. (b) The results from the replication sample. No significant differences in ratings of Vigor were seen between groups. (c) the recreated results from Hamidovic *et al* (2009). Rs12364283 A/A homozygotes had decreases in stop reaction times in response to amphetamine, while no such decrease was seen for the A/G + G/G group. (d) The results from the replication sample. The A/A group and A/G + G/G group displayed decreases in stop reaction time in response to amphetamine administration (5 and 10 mg). (e) The recreated results from Dlugos *et al* (2010) (*FAAH*). Rs229633 C/C homozygotes had increased ratings on the POMS Arousal scale in response to 10 mg amphetamine, as compared with the heterozygote group. (f) The results from the replication sample. There were no significant differences in ratings of Arousal between the three genotypes at 10 mg. $^{*}P<0.05$, $^{**}P<0.001$, $^{***}P<10^{-4}$.

Supplementary Figure S6A). *Post hoc* tests revealed that the C/C group had significantly higher ratings of Arousal in response to 10 mg amphetamine as compared with the C/T group ($P=0.003$); the C/C group also showed significantly reduced Fatigue ($P=0.005$). Additionally, significant Drug $\times$ Genotype interactions were found for rs3766246 and POMS Arousal ($P=0.013$) and POMS Fatigue ($P=0.009$). We obtained the same results in our repeat analysis. Participants in the C/C group reported higher ratings of Arousal and lower ratings of Fatigue when compared with participants in the C/T group ($P=0.009$ and $P=0.01$, respectively). Additionally, Dlugos *et al* found that the rs3766246–rs324420–rs2295633 CCC and TAT haplotypes were significantly associated with ratings of Fatigue at 10 mg, and we repeated the analysis and obtained the same results (CCC, $P=0.003$; TAT, $P=0.012$).

We conducted the same analysis used in the original publication in the replication sample ($N=173$). In the replication sample, the genotype groups for both SNPs did not differ on any measure (Figure 3f; Supplementary Figure S6B). Furthermore, neither the CCC nor the TAT haplotype was significantly associated with ratings of fatigue in the replication sample (CCC, $P=0.667$; TAT, $P=1.0$). Thus, we failed to replicate the results from the original sample.

COMT (Hamidovic *et al*, 2010a)

The original analysis of the catechol-*O*-methyltransferase gene (*COMT*) consisted of 161 Caucasian participants genotyped at the val158met polymorphism (rs4680). This SNP was analyzed in relation to subjective and behavioral responses to amphetamine administration (POMS

subscales, DSST; AUC) using 3×3 repeated-measures ANOVAs (Dose $\times$ Genotype). Paired-samples *t*-tests were used to assess the effect of drug on each genotype group when a significant Drug $\times$ Genotype interaction was found. Hamidovic *et al* (2010a) identified a significant Drug $\times$ Genotype interaction ($P = 0.008$) between scores of the DSST in response to amphetamine. We repeated this analysis and obtained the same result (Figure 4a). *Post hoc* analyses revealed that met/met carriers did not respond to amphetamine, while val/val carriers showed enhanced performance in the 10 mg and 20 mg drug sessions as compared with placebo (10 mg, $P = 5.4 \times 10^{-5}$; 20 mg, $P = 1.3 \times 10^{-4}$; Figure 4a). Val/met carriers showed an intermediate response to drug in the 20-mg session ($P = 0.002$).

We conducted the same analysis used in the original publication in the replication sample ($N = 176$), but did not identify any differences between the three genotype groups (Figure 4b), indicating a failure to replicate the results from the original sample.

SLC6A3 (Hamidovic *et al*, 2010b)

The original analysis of the dopamine transporter gene (*SLC6A3*) consisted of 152 Caucasian participants genotyped at four SNPs in *SLC6A3* (rs460000, rs3756450, rs37022, rs6869645). Due to low minor allele frequency, the minor allele homozygotes were pooled with the heterozygotes for all four SNPs. These SNPs were analyzed in relation to subjective effects and cognitive performance in response to amphetamine using 3×2 repeated-measures ANOVAs (Dose $\times$ Genotype). Paired-samples *t*-tests were used to assess the effect of drug on each genotype group when a significant Drug $\times$ Genotype interaction was found. Hamidovic *et al* (2010b) identified a significant Drug $\times$ Genotype interaction for the ARCI Amphetamine and ARCI MBG scales (AUC) and genotype at rs460000 ($P = 0.015$, $P = 0.025$, respectively). We repeated this analysis and obtained the same result (Supplementary Figure S7A; Figure 4c). *Post hoc* tests demonstrated that the C/C group had greater response to amphetamine when compared with the A/A + A/C group (ARCI Amphetamine placebo *vs* 20 mg $P = 3.1 \times 10^{-13}$ *vs* $P = 1.9 \times 10^{-7}$; ARCI MBG placebo *vs* 20 mg $P = 1.5 \times 10^{-11}$ *vs* $P = 2 \times 10^{-6}$).

We conducted the same analysis used in the original publication in the replication sample ($N = 169$). In the replication sample, there was a significant Drug $\times$ Genotype interaction between ratings on the ARCI Amphetamine subscale and genotype at rs460000; however, a *post hoc* one-way ANOVA revealed that this effect was driven by the placebo session ($P = 0.028$; Supplementary Figure S7B). There was no evidence of association with ARCI MBG, with all groups responding similarly across all sessions (Figure 4d), indicating a failure to replicate the results from the original sample.

OPRM1 (Dlugos *et al*, 2011)

The original analysis of the opioid receptor, mu 1 gene (*OPRM1*) consisted of 162 Caucasian participants genotyped at seven SNPs in *OPRM1* (rs1799971, rs510769, rs660756, rs1918760, rs2281617, rs1998220, rs1998220). These SNPs, along with seven haplotypes comprised of these SNPs, were analyzed in relation to the subjective response to amphetamine (ARCI subscales; PCS) using 3×3 repeated-measures ANOVAs/ANCOVAs (Dose $\times$ Genotype). *Post hoc* analyses consisted of one-way ANOVAs/ANCOVAs. Gender was used as a covariate in the analyses of ARCI BG, as it was seen to be associated with this subscale. Dlugos *et al* (2011) identified significant Drug $\times$ Genotype interactions between rs510769 and ARCI MBG ($P = 0.031$) and ARCI Amphetamine ($P = 0.019$), as well as between rs2281617 and ARCI MBG ($P = 0.01$) and ARCI BG ($P = 0.008$). We repeated this analysis and obtained the same result (Supplementary Figures S8A and C; Figure 4e; Supplementary Figure S8E). *Post hoc* tests revealed that these associations were specific to the 10 mg session. The rs510769 G/G group had increased ratings on the ARCI MBG scale as compared with the A/A group ($P = 0.02$; Supplementary Figure S8A), and the A/G and G/G groups had increased ARCI Amphetamine ratings as compared with the A/A group ($P = 0.005$, $P = 0.003$, respectively; Supplementary Figure S8C). The rs2281617 C/C group had increased ratings on the ARCI MBG and ARCI BG scales compared with the C/T + T/T group ($P = 3.4 \times 10^{-4}$, $P = 1.3 \times 10^{-4}$, respectively; Figure 4e; Supplementary Figure S8E).

Dlugos *et al*, also identified significant associations between the rs1799171–rs510769 AG and AA haplotypes and ratings on the ARCI Amphetamine scale, the rs1799171–rs510769 AA haplotype and ratings on the ARCI MBG scale, the rs1918760–rs2281617–rs1998220 ATA haplotype and ratings on the ARCI MBG scale, and the rs1918760–rs2281617–rs1998220 ATA and GCG haplotypes and ratings on the ARCI BG subscale. We repeated the analysis and obtained similar results (rs1799171–rs510769 AG and AA ARCI Amphetamine $P = 0.019$, $P = 0.005$; rs1799171–rs510769 AA and ARCI MBG $P = 0.031$; rs1918760–rs2281617–rs1998220 ATA and ARCI MBG $P = 0.01$; rs1918760–rs2281617–rs1998220 ATA and ARCI BG $P = 0.048$; rs1918760–rs2281617–rs1998220 GCG and ARCI BG $P = 0.069$).

We conducted the same analysis used in the original publication in the replication sample ($N = 171$). In the replication sample, the genotype groups for both SNPs did not show significant differences for any measures (Supplementary Figures S8B and D; Figure 4f; Supplementary Figure S8F). Furthermore, we failed to identify any associations with haplotypes in the replication sample (rs1799171–rs510769 AG and AA ARCI Amphetamine $P = 0.222$, $P = 0.190$; rs1799171–rs510769 AA and ARCI MBG $P = 0.590$; rs1918760–rs2281617–rs1998220 ATA and ARCI MBG $P = 0.233$; rs1918760–rs2281617–rs1998220 ATA and ARCI BG $P = 0.975$; rs1918760–rs2281617–rs1998220 GCG and ARCI BG $P = 0.159$). Taken together, these results reflect a failure to replicate the results from the original sample.

Population Stratification Analyses

The original sample of 99 participants, which was used in the analysis of *ADORA2A* (Hohoff *et al*, 2005), *SLC6A3* (Lott *et al*, 2005), *BDNF* (Flanagin *et al*, 2006), *SLC6A4* (Lott *et al*, 2006), *CSNK1E* (Veenstra-VanderWeele *et al*, 2006), and

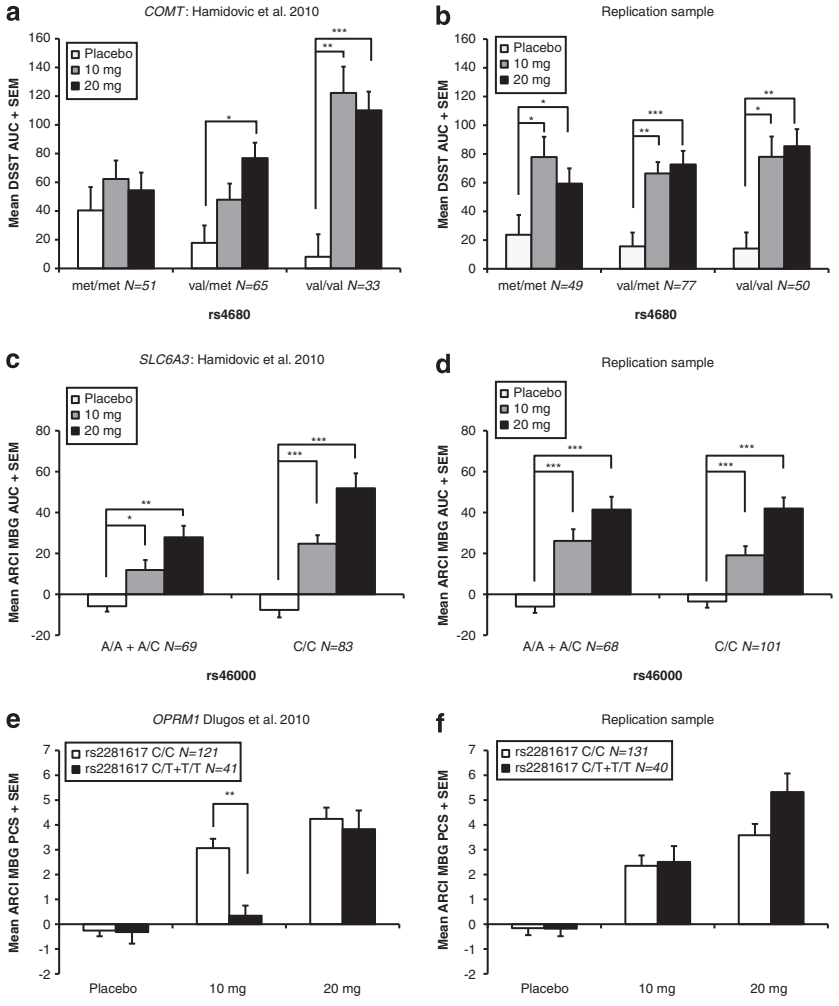

Figure 4 Replication results: *COMT*, *SLC6A3*, and *OPRM1*. (a) The recreated results from Hamidovic *et al* (2010a) (*COMT*). Val/met heterozygotes had increased performance on the DSST task in response to 20 mg amphetamine, while val/val homozygotes showed increased performance in response to both 10 and 20 mg amphetamine. No improvement was seen in the met/met group. (b) The results from the replication sample. Similar increases in performance were seen across all genotype groups. (c) The recreated results from Hamidovic *et al* (2010b) (*SLC6A3*). The rs46000 C/C group had increased subjective responding on the ARCI MBG scale in comparison to the A/A + A/C group. The results from the replication sample are found in (d); the C/C group and the A/A + A/C group showed very similar levels of subjective response to amphetamine. (e) The recreated results from Dlugos *et al* (2011) for rs2281617 (*OPRM1*). C/C homozygotes had increased ratings on the ARCI MBG subscale in response to 10 mg *d*-amphetamine as compared with the C/T + T/T group. (f) The results from the replication sample. The same increase was not seen in the C/C group as compared with the C/T + T/T group. $*P<0.05$, $**P<0.001$, $***P<10^{-4}$.

SLC6A2 (Dlugos *et al*, 2007), included 41 participants who self-reported being non-Caucasians (subsequent studies of *SLC6A2*, *DRD2*, *FAAH*, *COMT*, and *OPRM1* excluded non-Caucasians). We used genome-wide SNP data (Hart *et al*, 2012) to evaluate whether population stratification contributed to the associations identified using the non-Caucasian participants. Because nine samples were not available for genotyping with the Affymetrix 6.0 microarray, this analysis required us to exclude those nine participants. We report the initial *P*-values ('Original'), the effect of removing those nine individuals ('Original—9 participants') and any additional effect of using the ancestry PCs as covariates ('Original – 9 participants + PCs') in Table 3. Adjustment for ancestry principal components appeared to have little impact on the initial results. When we included the ancestry PCs as covariates, some associations became slightly stronger (eg, *SLC6A3* and ARCI LSD), while others became slightly weaker (eg, *CSNK1E* and ARCI MBG; Table 3). Taken together, these results demonstrate that none of the original associations appears to be primarily due to population stratification.

DISCUSSION

Our results show an unexpectedly widespread failure to replicate our previously published findings. This study is striking because we were attempting to replicate apparently robust findings related to well-studied candidate genes. We used a relatively large number of new participants for the replication, and their data were collected and analyzed using identical procedures. Thus, our study did not suffer from the heterogeneity in phenotyping procedures implicated in previous failures to replicate other candidate gene

Table 3 Association *P*-values from Tests With or Without Incorporation of Ancestry Principal Components (PCs) as Covariates in Analyses Involving Non-Caucasians

Gene	Original	Original—9 participants	Original—9 participants + PCs
ADORA2A	POMS Anxiety: 0.041 ($N=98$)	POMS Anxiety: 0.107 ($N=89$)	POMS Anxiety: 0.114 ($N=89$)
SLC6A3	DEQ Feel: 0.006 ($N=84$)	DEQ Feel: 0.006 ($N=84$)	DEQ Feel: 0.013 ($N=84$)
	ARCI LSD: 0.007 ($N=86$)	ARCI LSD: 0.007 ($N=86$)	ARCI LSD: 0.005 ($N=86$)
	Diastolic BP: 0.037 ($N=95$)	Diastolic BP: 0.022 ($N=90$)	Diastolic BP: 0.068 ($N=90$)
BDNF	POMS Arousal: 0.01 ($N=94$)	POMS Arousal: 0.013 ($N=89$)	POMS Arousal: 0.005 ($N=89$)
	ARCI BG: 0.023 ($N=94$)	ARCI BG: 0.015 ($N=89$)	ARCI BG: 0.018 ($N=89$)
	Heart rate: 0.023 ($N=94$)	Heart rate: 0.025 ($N=89$)	Heart rate: 0.012 ($N=89$)
SLC6A4	ARCI MBG: 0.046 ($N=98$)	ARCI MBG: 0.071 ($N=90$)	ARCI MBG: 0.064 ($N=90$)
CSNK1E	DEQ Feel: 0.038 ($N=88$)	DEQ Feel: 0.064 ($N=85$)	DEQ Feel: 0.053 ($N=85$)
	ARCI MBG: 0.008 ($N=89$)	ARCI MBG: 0.024 ($N=87$)	ARCI MBG: 0.068 ($N=87$)
SLC6A2	POMS Pos Mood: 0.012 ($N=90$)	POMS Pos Mood: 0.012 ($N=90$)	POMS Pos Mood: 0.008 ($N=90$)
	POMS Elation: 0.003 ($N=90$)	POMS Elation: 0.003 ($N=90$)	POMS Elation: 0.005 ($N=90$)

Nine samples from the original analysis were not available for genotyping with the Affymetrix 6.0 microarray and therefore did not have ancestry PCs. Sample size is given in parentheses.

studies (Ho *et al*, 2010; Mathieson *et al*, 2012). The failure of our associations to replicate suggests that most or all of our original results were false positives.

One possible cause of these false positives could have been that that six of our original studies included 41 non-Caucasian participants (Table 3). To address this concern, we repeated the original analyses with the addition of ancestry PCs as covariates; there were no major differences. Therefore, while population stratification can sometimes lead to false positive results, it does not appear that the inclusion of non-Caucasians significantly contributed to the observed failure to replicate our previously published results.

It is worth considering whether we should have viewed our original results with greater skepticism. Genome-wide association studies (GWAS) of a wide variety of phenotypes suggest that the effects of individual alleles are very small, such that the modestly sized samples typically used in candidate-gene studies such as ours would be severely under-powered (McCarthy *et al*, 2008). Both the original and the replication samples were too small to detect alleles with the small effect sizes seen in GWAS. Our original studies suggested that we were detecting alleles that contributed $\sim$5% of the total phenotypic variance we reported. Given that there are millions of polymorphisms in the human genome, such large effects might have aroused greater scrutiny, but we were reassured by the commonly held belief that polymorphisms in our candidate genes represented a privileged subset of polymorphisms and by the notion that intermediate phenotypes might have a simpler genetic architecture. We were not alone—many other candidate gene studies have and continue to report similarly large effect sizes and to espouse similar beliefs.

Three of our previously reported associations were due to the *lack* of a drug effect in a particular genotype group; in all cases, the rare homozygote groups did not show a significant drug response, which could reflect a lack of power rather than a true lack of response. For example, the 3′ UTR VNTR polymorphism in *SLC6A3* was associated with ratings of DEQ Feel in Lott *et al* (2005), but in the original analysis the minor allele 9/9 genotype group ($N=7$) did not show response to amphetamine, while the heterozygote ($N=33$) and major allele 10/10 ($N=52$) groups did. In the replication sample, the 9/9 group ($N=24$), like the 9/10 and 10/10 groups, showed a significant drug response. Similarly, a lack of drug effect in rare allele homozygote groups contributed to the associations in Flanagin *et al* (2006) (*BDNF*) and Hamidovic *et al* (2010a) (*COMT*); these results reflect poor power to detect the effect of amphetamine due to a small number of rare allele homozygotes. This phenomenon has been noted in other candidate gene studies of the *SLC6A3* 3′ UTR where a lack of effect was observed in the 9/9 genotype group (Joober *et al*, 2007; Stein *et al*, 2005), suggesting that this may be a widespread problem. The fundamental issue is that small genotype groups may not show a response to treatment due to a lack of power. One potentially valuable strategy to avoid this problem is prospective genotyping, which allows for more balanced genotype groups and is thus helpful when evaluating rare alleles.

Two related problems that are common among candidate gene studies like ours are insufficient correction for multiple testing (both within and across studies) and publication bias. Although several of our previous reports applied corrections for the number of tests performed *within* that publication, others did not. Furthermore, we never corrected for all comparisons performed *across* all 12 studies. Similar failures to fully correct for multiple testing are common in the candidate gene literature, where large data sets are often repeatedly analyzed. If we corrected for all 322 primary tests performed in this study, the Bonferroni-corrected significance threshold would be 0.00015. While this *P*-value is overly stringent because both SNPs and phenotypes are inter-correlated, it gives some sense for the cumulative burden of multiple testing across all 12 studies. Multiple testing across studies is more problematic than multiple testing within studies, as it is often not readily apparent that a data set has been analyzed

repeatedly from the reading of one study. Better standards for reporting prior analyses of a given data set might be helpful, but in the end running more tests will inevitably inflate the number of false positives, whether the tests use one data set repeatedly or many separate data sets. Thus standards that tend to preclude multiple analyses of the same data set are too simplistic to fully address this problem.

The problem of publication bias, which is the tendency to preferentially publish significant results and to repress non-significant ones, is related to the failure to correct for multiple testing because the true number of hypotheses tested in a given data set is concealed. It has been argued that publication bias against non-significant results contributes to non-replication of candidate gene associations (Bosker *et al*, 2011; Munafò *et al*, 2007). Our original results reflect a minor degree of publication bias: in our early investigations we performed preliminary analyses on a small number of genes that did not yield significant results and thus we did not publish them. This increase in multiple testing was not taken into account when determining significance thresholds. Similarly, we sometimes considered several alternative methods for calculating phenotypes (eg, peak change score summarization *vs* area under the curve, which tend to be highly but incompletely correlated). It seems very likely that the candidate gene literature frequently reflects this sort of publication bias, which represents a special case of uncorrected multiple testing. Proper correction for multiple phenotypes is a concern and a source of debate for multidimensional phenotypes, such as brain imaging (Poldrack and Mumford, 2009; Bennett *et al*, 2011).

One feature of our studies was the use of subjective drug effects as outcome measures. While we initially regarded the use of subjective drug effects as a strength, it may be that other phenotypes provide a more sensitive indicator of drug response. Although subjective drug effects are dose and time dependent and provide a unique, face-valid indicator of the drug's effect on behavior, they are also highly variable within and across participants, and are subject to the biases present in any self-report measure. Instead, measures such as functional magnetic resonance imaging (fMRI) may provide a more precise and objective index of biological response to a drug. One example of the success of fMRI phenotypes has been the association of the amygdala response to threat stimuli and the 5-HTTLPR polymorphism, which was initially suggested to account for 10% of the phenotypic variance (Munafò *et al*, 2008) but was later determined to account for $\sim$1% (Murphy *et al*, 2012). Other examples of potentially promising phenotypes include alcohol-induced flushing (Macgregor *et al*, 2009; Wall *et al*, 2005), nicotine metabolism (Benowitz *et al*, 2003; Lerman *et al*, 2006), and electroencephalography (Hodgkinson *et al*, 2010). Ultimately, the optimal phenotype for any particular scientific or clinical question depends on the sensitivity and selectivity of the measure and practical issues such as cost and throughput.

Using our results as an example, we demonstrate that a rigorously assessed and biologically based intermediate phenotype has the potential to yield false positives, in part because of the common practices used in candidate gene studies. How can this problem be addressed in the future? Clearly more stringent thresholds for significance that better address multiple testing within and between studies are important. Because an initial study is best for hypothesis generation, replication studies, such as this one, are also essential. More stringent standards will require correspondingly larger samples. Although many journals have added requirements for replication (Anonymous, 2005; Barsh *et al*, 2012; Hewitt 2012; http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1552-485X/homepage/ForAuthors.html; http://www.blackwellpublishing.com/pdf/G2B-Association-Studies.pdf), replication is not infallible (Sullivan, 2007).

Candidate gene studies with samples of several hundred or less subjects are only adequately powered to detect alleles with effects that are significantly larger than those observed in most GWAS. Thus, these studies are only productive if one assumes that an intermediate phenotype is fundamentally different from a disease trait. Our study does not support this hypothesis, although it just one example. If the genetic architecture of intermediate phenotypes is indeed similar to disease traits, very large samples will be needed to achieve sufficient power.

One key factor for a successful candidate gene study is to have strong prior information that the polymorphism being examined is likely to be a true positive. Whereas 'traditional' candidate gene studies such as ours have focused on heavily studied genes (sometimes with a specific focus on coding SNPs, eg Flanagin *et al*, 2006; Hamidovic *et al*, 2010a; Dlugos *et al*, 2011), a more recent trend is to focus on the SNPs that have experimentally validated effects on gene expression; such SNPs are termed expression quantitative trait loci (eQTL). Recent studies have shown that SNPs associated with complex traits are enriched for eQTLs (Schadt *et al*, 2008; Nicolae *et al*, 2010; Fehrmann *et al*, 2011; Gamazon *et al*, 2012). Similarly, recent GWAS studies have begun to provide unambiguous associations between SNPs and disease traits (Furberg *et al*, 2010; Ripke *et al*, 2011); these SNPs are likely to be the subject of the next wave of candidate gene studies. While focusing on polymorphisms that have known biological effects can only improve candidate gene studies, the fundamental question remains: is it realistic to assume that the effect of these SNPs will be large enough to allow for detection when examining intermediate phenotypes with only modestly sized samples?

In conclusion, in an effort to examine the validity and replicability of our previous work, we performed a replication study of 12 of our previously published candidate gene association studies. We were motivated to perform this replication study because we believed that we had an ideal sample to explore replication in a broad range of different candidate genes. *We failed to replicate any of our previously published results, suggesting that our previously published findings were likely false positives*. More broadly, our results should instill caution in other investigators who, in some cases inspired by our previous publications, have undertaken similarly designed and powered studies. The final judgment about the usefulness of intermediate phenotypes will depend on the results from many studies. Our experience provides one example in which a promising intermediate phenotype did not perform as expected. We conclude that future candidate gene studies focused on intermediate phenotypes similar to ours should

strongly consider the possibility that effect sizes may be similar to those observed in GWAS.

ACKNOWLEDGEMENTS

This work was supported by NIH Grants DA007255 (ABH), DA02812 (HdW), and DA021336 and DA024845 (AAP). We thank Barbara E Engelhardt for providing imputed SNP genotypes and Margaret C Wardle for organization and preprocessing of the phenotype data.

DISCLOSURE

HdW has received a research grant from Unilever for a project unrelated to this study. ABH and AAP declare no potential conflict of interest.

REFERENCES

Alexander RC, Wright R, Freed W (1996). Quantitative trait loci contributing to phencyclidine-induced and amphetamine-induced locomotor behavior in inbred mice. *Neuropsychopharmacology* **15**: 484–490.

American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. Editorial Policy on Association Studies http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1552-485X/homepage/ForAuthors.html.

Anonymous (2005). Framework for a fully powered risk engine. *Nat Genet* **37**: 1153.

Barsh GS, Copenhaver GP, Gibson G, Williams SM (2012). Guidelines for genome-wide association studies. *Plos Genet* **8**: e1002812.

Bennett C, Baird AA, Miller MB, Wolford GL (2011). Neural correlates of interspecies perspective taking in the post-Mortem Atlantic Salmon: an argument for proper multiple comparisons correction. *J Serendipitous Unexpected Results* **1**: 1–5.

Benowitz NL, Pomerleau OF, Pomerleau CS, Jacob P (2003). Nicotine metabolite ratio as a predictor of cigarette consumption. *Nicotine Tob Res* **5**: 621–624.

Bosker FJ, Hartman CA, Nolte IM, Prins BP, Terpstra P, Posthuma D *et al* (2011). Poor replication of candidate genes for major depressive disorder using genome-wide association data. *Mol Psychiatry* **16**: 516–532.

Chait L, Fischman MW, Schuster CR (1985). 'Hangover' effects the morning after marijuana smoking. *Drug Alcohol Depend* **15**: 229–238.

Crabbe J, Jarvik L, Liston E, Jenden D (1983). Behavioral responses to amphetamines in identical twins. *Acta Genet Med Gemellol (Roma)* **32**: 139–149.

Daly AK (2010). Genome-wide association studies in pharmacogenomics. *Nat Rev Genet* **11**: 241–246.

Dlugos A, Freitag C, Hohoff C, McDonald J, Cook E, Deckert J *et al* (2007). Norepinephrine transporter gene variation modulates acute response to D-amphetamine. *Biol Psychiatry* **61**: 1296–1305.

Dlugos AM, Hamidovic A, Hodgkinson C, Shen PH, Goldman D, Palmer AA *et al* (2011). OPRM1 gene variants modulate amphetamine-induced euphoria in humans. *Genes Brain Behav* **10**: 199–209.

Dlugos AM, Hamidovic A, Hodgkinson CA, Goldman D, Palmer AA, de Wit H (2010). More aroused, less fatigued: fatty acid amide hydrolase gene polymorphisms influence acute response to amphetamine. *Neuropsychopharmacology* **35**: 613–622.

Dlugos AM, Hamidovic A, Palmer AA, Wit H (2009). Further evidence of association between amphetamine response and SLC6A2 gene variants. *Psychopharmacology (Berl)* **206**: 501–511.

Durbin R, Abecasis G, Altshuler D, Auton A, Brooks L, Gibbs R *et al* (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Fehrmann RSN, Jansen RC, Veldink JH, Westra H-J, Arends D, Bonder MJ *et al* (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *Plos Genet* **7**: e1002197.

Flanagin B, Cook Jr E, de Wit H (2006). An association study of the brain-derived neurotrophic factor Val66Met polymorphism and amphetamine response. *Am J Med Genet B* **141**: 576–583.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA *et al* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D *et al* (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* **42**: 441–447.

Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ *et al* (2012). Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatryhttp://www.blackwell-publishing.com/pdf/G2B-Association-Studies.pdf.*

Goldman D, Ducci F (2007). Deconstruction of vulnerability to complex diseases: enhanced effect sizes and power of intermediate phenotypes. *ScientificWorldJournal* **7**: 124–130.

Gottesman II, Gould TD (2003). The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**: 636–645.

Grisel JE, Belknap JK, O'Toole LA, Helms ML, Wenger CD, Crabbe JC (1997). Quantitative trait loci affecting methamphetamine responses in BXD recombinant inbred mouse strains. *J Neurosci* **17**: 745–754.

Hamidovic A, Dlugos A, Palmer A, de Wit H (2010a). Catechol-O-methyltransferase val158met genotype modulates sustained attention in both the drug-free state and in response to amphetamine. *Psychiatr Genet* **20**: 85.

Hamidovic A, Dlugos A, Palmer AA, de Wit H (2010b). Polymorphisms in dopamine transporter (SLC6A3) are associated with stimulant effects of d-amphetamine: an exploratory pharmacogenetic study using healthy volunteers. *Behav Genet* **40**: 255–261.

Hamidovic A, Dlugos A, Skol A, Palmer AA, de Wit H (2009). Evaluation of genetic variability in the dopamine receptor D2 in relation to behavioral inhibition and impulsivity/sensation seeking: an exploratory study with d-amphetamine in healthy participants. *Exp Clin Psychopharmacol* **17**: 374–383.

Hart AB, Engelhardt BE, Wardle MC, Sokoloff G, Stephens M, de Wit H *et al* (2012). Genome-wide association study of d-amphetamine response in healthy volunteers identifies putative associations, including cadherin 13 (*CDH13*). *PLoS ONE* **7**: e42646.

Hewitt JK (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav Genet* **42**: 1–2.

Ho MK, Goldman D, Heinz A, Kaprio J, Kreek MJ, Li MD *et al* (2010). Breaking barriers in the genomics and pharmacogenetics of drug addiction. *Clin Pharmacol Ther* **88**: 779–791.

Hodgkinson CA, Enoch M-A, Srivastava V, Cummins-Oman JS, Ferrier C, Iarikova P *et al* (2010). Genome-wide association identifies candidate genes that influence the human electroencephalogram. *Proc Natl Acad Sci USA* **107**: 8695–8700.

Hodgkinson CA, Yuan Q, Xu K, Shen P-H, Heinz E, Lobos EA *et al* (2008). Addictions biology: haplotype-based analysis for 130 candidate genes on a single array. *Alcohol Alcohol* **43**: 505–515.

Hohoff C, Mcdonald JM, Baune BT, Cook EH, Deckert J, de Wit H (2005). Interindividual variation in anxiety response to amphetamine: possible role for adenosine A 2Areceptor gene variants. *Am J Med Genet* **139B**: 42–44.

Howie BN, Donnelly P, Marchini J (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *Plos Genet* **5**: e1000529.

Johanson C, Uhlenhuth E (1980). Drug preference and mood in humans: diazepam. *Psychopharmacology (Berl)* **71**: 269–273.

Joober R, Grizenko N, Sengupta S, Amor LB, Schmitz N, Schwartz G *et al* (2007). Dopamine transporter 3'-UTR VNTR genotype and ADHD: a pharmaco-behavioural genetic study with methylphenidate. *Neuropsychopharmacology* **32**: 1370–1376.

Kamens HM, Burkhart-Kasch S, McKinnon CS, Li N, Reed C, Phillips TJ (2005). Sensitivity to psychostimulants in mice bred for high and low stimulation to methamphetamine. *Genes Brain Behav* **4**: 110–125.

Lerman C, Tyndale R, Patterson F, Wileyto EP, Shields PG, Pinto A *et al* (2006). Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation. *Clin Pharmacol Ther* **79**: 600–608.

Logan G, Cowan W, Davis K (1984). On the ability to inhibit simple and choice reaction time responses: a model and a method. *J Exp Psychol* **10**: 276–291.

Lott D, Kim S, Cook E, de Wit H (2005). Dopamine transporter gene associated with diminished subjective response to amphetamine. *Neuropsychopharmacology* **30**: 602–609.

Lott D, Kim S, Cook E Jr, de Wit H (2006). Serotonin transporter genotype and acute subjective response to amphetamine. *Amer J Addiction* **15**: 327–335.

Macgregor S, Lind PA, Bucholz KK, Hansell NK, Madden PAF, Richter MM *et al* (2009). Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. *Hum Mol Genet* **18**: 580–593.

Martin W, Sloan J, Sapira J, Jasinski D (1971). Physiologic, subjective, and behavioral effects of amphetamine, methamphetamine, ephedrine, phenmetrazine, and methylphenidate in man. *Clin Pharmacol Ther* **12**: 245–258.

Mathieson I, Munafò MR, Flint J (2012). Meta-analysis indicates that common variants at the DISC1 locus are not associated with schizophrenia. *Mol Psychiatry* **17**: 634–641.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA *et al* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**: 356–369.

Munafò MR, Brown SM, Hariri AR (2008). Serotonin transporter (5-HTTLPR) genotype and amygdala activation: a meta-analysis. *Biol Psychiatry* **63**: 852–857.

Munafò MR, Matheson IJ, Flint J (2007). Association of the DRD2 gene Taq1A polymorphism and alcoholism: a meta-analysis of case-control studies and evidence of publication bias. *Mol Psychiatry* **12**: 454–461.

Murphy SE, Norbury R, Godlewska BR, Cowen PJ, Mannie ZM, Harmer CJ *et al* (2012). The effect of the serotonin transporter polymorphism (5-HTTLPR) on amygdala function: a meta-analysis. *Mol Psychiatry* (e-pub ahead of print).

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *Plos Genet* **6**: e1000888.

Nurnberger J, Gershon ES, Simmons S, Ebert M, Kessler L, Dibble E *et al* (1982). Behavioral, biochemical and neuroendocrine responses to amphetamine in normal twins and 'well-state' bipolar patients. *Psychoneuroendocrinology* **7**: 163–176.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *Plos Genet* **2**: e190.

Poldrack RA, Mumford JA (2009). Independence in ROI analysis: where is the voodoo? *Soc Cogn Affect Neurosci* **4**: 208–213.

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA *et al* (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**: 969–976.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY *et al* (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**: e107.

Stein MA, Waldman ID, Sarampote CS, Seymour KE, Robb AS, Conlon C *et al* (2005). Dopamine transporter genotype and methylphenidate dose response in children with ADHD. *Neuropsychopharmacology* 1–9.

Sullivan PF (2007). Spurious genetic associations. *Biol Psychiatry* **61**: 1121–1126.

Veenstra-VanderWeele J, Qaadir A, Palmer AA, Cook EH, de Wit H (2006). Association between the Casein Kinase 1 Epsilon gene region and subjective response to D-amphetamine. *Neuropsychopharmacology* **31**: 1056–1063.

Wall TL, Shea SH, Luczak SE, Cook TAR, Carr LG (2005). Genetic associations of alcohol dehydrogenase with alcohol use disorders and endophenotypes in white college students. *J Abnorm Psychol* **114**: 456–465.

Wechsler D (1958). The measurement and appraisal of adult intelligence. *J Med Educ* **33**: 706.

White TL, Justice AJH, de Wit H (2002). Differential subjective effects of d-amphetamine by gender, hormone levels and menstrual cycle phase. *Pharmacology, Biochemistry and Behavior* **73**: 729–741.

Zombeck JA, Swearingen SP, Rhodes JS (2010). Acute locomotor responses to cocaine in adolescents vs adults from four divergent inbred mouse strains. *Genes Brain Behav* **9**: 892–898.

Supplementary Information accompanies the paper on the Neuropsychopharmacology website (http://www.nature.com/npp)