HAL
open science

Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia

Danielle Posthuma, Esther S Lips, L Niels Cornelisse, Ruud F Toonen, Josine L Min, Christina Hultman, International Schizophrenia Consortium, Peter A Holmans, Michael C. O'Donovan, Shaun Purcell, et al.

To cite this version:

Danielle Posthuma, Esther S Lips, L Niels Cornelisse, Ruud F Toonen, Josine L Min, et al.. Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. Molecular Psychiatry, 2011, 10.1038/mp.2011.117 . hal-00680704

HAL Id: hal-00680704

https://hal.science/hal-00680704v1

Submitted on 20 Mar 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Running title: *Synaptic gene groups and Schizophrenia*

Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia

Esther S Lips[1], L Niels Cornelisse[1], Ruud F Toonen[1], Josine L Min[1], Christina M Hultman[2,3], the International Schizophrenia Consortium[#], Peter A. Holmans[4], Michael C. O'Donovan[4] Shaun M. Purcell[5,6,7,8], August B Smit[9], Matthijs Verhage[1], Patrick F Sullivan[10], Peter M Visscher[11], Danielle Posthuma[1,12]

[1] Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam VU University, The Netherlands
[2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden,
[3] Department of Neuroscience, Psychiatry, Ullerråker, Uppsala University, Uppsala, Sweden,
[4] School of Medicine, Department of Psychological Medicine, School of Medicine, Cardiff University, Cardiff, United Kingdom,
[5] Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America,
[6] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America,
[7] Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America,
[8] Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America,
[9] Molecular & Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam VU University, The Netherlands,
[10] Department of Genetics, University of North Carolina, Chapel Hill, United States of America,
[11] Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, Brisbane, Australia,
[12] Department of Medical Genomics, VU Medical Center, Neuroscience Campus Amsterdam, The Netherlands,
Please see Acknowledgements for consortium authorship

Correspondence to: Danielle Posthuma, Center for Neurogenomics and Cognitive Research (CNCR), De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands Tel: +31 20 598 2823, Fax: +31 20 598 6926, E-mail: danielle.posthuma@cncr.vu.nl

Summary

Schizophrenia is a highly heritable disorder with a polygenic pattern of inheritance and a population prevalence around 1 percent. Previous studies have implicated synaptic dysfunction in schizophrenia. We tested the accumulated association of genetic variants in expert curated synaptic gene groups with schizophrenia in 4,673 cases and 4,965 healthy controls, using functional gene group analysis. Identifying groups of genes with similar cellular function rather than genes in isolation may have clinical implications for finding additional drug targets. We found that a group of 1026 synaptic genes was significantly associated with the risk of schizophrenia ($P < 7.6x10^{-11}$) and more strongly associated than 100 randomly drawn, matched control groups of genetic variants ($P < .01$). Subsequent analysis of synaptic sub-groups suggested that the strongest association signals are derived from three synaptic gene groups: intracellular signal transduction ($P = 2.0x10^{-4}$), excitability ($P = 9.0x10^{-4}$) and cell adhesion and trans-synaptic signaling ($P = 2.4x10^{-3}$). These results are consistent with a role of synaptic dysfunction in schizophrenia and imply that impaired intracellular signal transduction in synapses, synaptic excitability and cell adhesion and trans-synaptic signaling play a role in the pathology of schizophrenia.

Keywords: GWAS, ISC, GAIN, gene group analysis, synapse, schizophrenia, genome-wide association

Introduction

Schizophrenia is a chronic and debilitating brain disorder that affects about 1 percent of the population[1]. It is characterized by delusional beliefs, hallucinations, disordered speech and deficits in emotional and social behavior[e.g.2], and is highly familial with heritability estimates of 81%[3].

GWA studies have explained only a small amount of genetic variance in schizophrenia and are limited in power due to the many tests performed and do not necessarily lead to knowledge about molecular mechanisms of a clinical trait. In addition, the strongest associated variants - when part of a pathway - might not be the best drug target for therapeutic intervention and identifying variants in the same cellular pathway or functionally related gene group may help in finding additional drug targets. Recent genome-wide association studies for schizophrenia have implicated the major histocompatibility complex on 6p21.2-22.1, neurogranin (*NRGN*), and transcription factor 4 (*TCF4*) [4,5,6]. In addition, they have provided molecular genetic evidence for a substantial polygenic component, implicating a large number of SNPs of very small effect in the etiology of schizophrenia[4]. Some of these exceed genome-wide significance, but the currently available sample sizes are insufficient to detect these effects[7]. Therefore, SNPs in the $5x10^{-8}$ -$1x10^{-6}$ band are a mix of SNPs, some of true effect and some that are false positives. Exactly what the mix of true/false positives is, is currently unknown. Gene set or pathway analysis involves testing for the combined effect of multiple SNPs, which individually may have a very small effect that does not reach significance. By using a competitive testing scheme, associations of gene sets with a disease are corrected for false positives.

It seems likely that the substantial polygenic component involves SNPs that are not distributed randomly across the genome but are distributed across genes that share a common biological function or pathway[8,9,10,11]. Recent pathway analyses provided evidence for the importance of the cell adhesion molecule (CAM) pathway in schizophrenia[12], as well as the glutamate metabolism, TGF-beta signaling and TNFR1 pathways[10].

Pathway analysis is predicated upon accurate pathway definitions and validated assignment of genes to pathways. However, many of the available databases used for pathway definitions are not optimally annotated and the same pathways can be differentially defined across databases. Classically defined pathways are usually not independent, as the same genes, especially the endpoints, are often active in different pathways. Consequently, genetic variation that affects the expression or function of genes in different pathways may have similar consequences, have similar impact on pathogenesis and show similar disease association. Genes may also be grouped according to shared cellular function[e.g.11]. Such 'functional gene grouping' goes across the traditionally defined biological pathways as it groups genes based on similar cellular function, and not based on a cascade of induced events, as in biological pathways. We recently proposed such a functional gene grouping approach to test for the combined effect of genetic variants in genes with shared cellular function in the synapse using a manually curated database of gene function based on both experimentation and data-mining[11]. Using this approach it was found that one relay element involved in many pathways (G-proteins) was associated with cognitive traits, a strong association, which had remained unnoticed by traditional single-marker analysis[11].

Numerous statistical methods are available to evaluate the enrichment of selected pathways or functional gene groups for selected traits, including e.g. the Gene Set Enrichment Analysis (GSEA)[13,14], testing for overrepresentation of categories of genes[9], the SNP ratio test[12], hypergeometric tests[e.g.10], or the Σ-log(P) method combined with permutation[11]. Most of these methods correct for linkage disequilibrium (LD) between SNPs, the number of SNPs per gene, gene size and multiple testing of independent pathways. Permutation is generally used to determine how likely a given result is if the null hypothesis of no association is true. However, the more genes are present in the defined gene-group the more likely it becomes to observe smaller P-values. In addition, generally these methods do not test how unique a certain result is given the polygenic nature of many studied traits. The latter involves testing whether a given pathway is significantly

associated with a trait because it a) includes a lot of genes and the trait is polygenic in nature, or b) because of the biological function of the pathway. This can be resolved by testing for association of matched-control gene-groups in comparison with the targeted gene groups or pathways.

The purpose of the current study is to apply a functional gene group approach to detect well-annotated functional gene groups that are important to the risk of schizophrenia. The synaptic hypothesis of schizophreni[15,16,17,18] is one of the leading hypotheses in the field of schizophrenia. Recent genetic findings underscore the importance of synaptic dysfunction in schizophrenia[6,12]. Therefore, we formally tested whether the group of genes involved in pre- and postsynaptic functioning is related to schizophrenia, and whether this group is more strongly related than randomly drawn matched sets of genes, using a "competitive" control method[19]. Apart from testing all synaptic genes for an association with the risk of schizophrenia, we also tested 17 subgroups of synaptic functioning, defined based on data-mining and experimentation[11]. We used the data genotyped within the International Schizophrenia Consortium (ISC) case-control sample[4] and the Genetic Association Information Network (GAIN) schizophrenia dataset.

Materials and Methods

Participants and genotyping

The ISC case-control sample includes 3,322 cases and 3,587 controls (European ancestry), derived from seven different collection sites and is described in detail elsewhere[4]. Subjects were genotyped on Affymetrix 5.0 or 6.0 SNP arrays, and 300,523 SNPs passed quality control for the ISC_affy5 sample (3353 subjects) and 717,126 SNPs for the ISC_affy6 sample (3556 subjects).

The GAIN schizophrenia case-control sample has been described elsewhere[5,20] and has been downloaded from dbGAP (phs000021.v2.p1). Briefly, this sample included 1351 cases and 1378 controls of European ancestry. All individuals were genotyped on the Affymetrix 6.0 array, and 727,872 SNPs were available for analysis. The Quality Control (QC) procedures followed those

described in Shi et al[5]. The GAIN and two ISC (affy5 and affy6) are independent and non-overlapping, and together contain 9638 individuals (4673 cases/4965 controls) of European ancestry.

Defining functional gene groups

Synaptic functional gene group definition was based on cellular function as determined by previous protein identification and data mining for synaptic genes and gene function[11]. Genes were considered 'synaptic' based on proteomic analysis of synaptic preparations[21]. In case of presynaptic genes, an additional expert curation was performed because only few analyses of highly purified preparations are currently available for the presynaptic proteome, except synaptic vesicles[22]. Hence, presynaptic genes not covered by[22] were manually curated using published functional data and a cumulative scoring paradigm with the following set of weighted criteria: null mutation produces a synaptic phenotype; activation of the gene product (e.g. receptor) or blockade thereof directly modulates synaptic function; immuno-electronmicroscopy detects gene product in the synapse. More than 500 PubMed entries were manually screened. Although this approach introduces a bias towards well-studied genes, this is inherent to creating functional gene groups, as functional grouping is by definition limited to those genes for which functional data is available. Synaptic genes were subdivided into seventeen functional groups based on shared cellular function (a full listing of genes assigned to functional groups is provided in the supplementary material Table S4).

SNP assignment

All SNPs that survived QC in the ISC and GAIN samples were mapped to genes on the basis of NCBI human assembly build 36.3 and dbSNP release 129 (following[9]). For the definition of the gene boundaries we downloaded the 'seq_gene.md' file from NCBI's ftp website. From this list of records we deleted genes coded as *pseudo* in the column 'feature_type'. Subsequently, we selected

the records with *gene* as 'feature_type' and *reference* as 'group_label'. For these records, we assigned SNPs to genes when annotated between 'chr_start' (transcription start site, TSS) and 'chr_stop' (transcription stop site, TES).

Association analysis

SNP association analyses were carried out using additive models of allele counts. For the ISC dataset, a correction for clustering within stratum (collection site) was performed[4]. Cochran-Mantel-Haenszel (CMH) tests implemented in PLINK were used for the association analyses. All analyses were carried out separately for the ISC_affy5, ISC_affy6 and the GAIN datasets. Empirical *P*-values from the three datasets were combined by Stouffer's weighted Z-transform method[23] to obtain an overall *P*-value.

Evaluating the combined effect of all SNPs in a functional gene group: the Σ-log(P) method

We summed the logarithm of the reciprocal of the P-values (denoted as Σ-log(P) method[24,25] as previously applied to gene group analysis[11] to determine the significance of the combined effect of SNPs annotated to genes in a functional group. The Σ-log(P) method combines P-values from association analyses within a group of genetic variants, then calculates the $-\log_{10}$ of each P-value, and sums over all P-values in a group to obtain the Σ-log(P) test statistic. To allow unbiased interpretation of the Σ-log(P) test statistic, 10,000 permutations were conducted which are implicitly conditional on LD, sample size, gene size, the number of SNPs per gene and the number of genes per group, by permuting affection status over genotypes. With this permutation procedure, only the relation between any genetic variant and affection status was disconnected, whereas LD structure was kept intact. In addition, each group of genetic variants included the same (numbers of) SNPs and genes and had the same sample size as the original dataset. For each permutation, we obtained the Σ-log(P) for each functional group and then compared the observed Σ-log(P) of a group to the

empirical P-value distribution by calculating the proportion of Σ-log(P) in the permuted datasets that was higher than the observed Σ-log(P).

Controlling for known polygenic effect on schizophrenia

The permutation approach described above provides information on how likely a given value of the combined effect of all SNPs in a group of genes is under the null hypothesis of no association of any SNP included in the functional group with the risk of schizophrenia (i.e. self-contained testing)[19]. We additionally applied matched-control methods (i.e. competitive testing) which allow to test whether randomly drawn groups of SNPs/genes would provide an equally or more significant (combined) empirical P-value as compared to the combined P-value from the group of synaptic genes. We created control gene-groups matched for the number of genes (method 1) and groups that were matched for the effective number of SNPs, which could be drawn from all genic and non-genic SNPs (method 2), from genic SNPs only (method 3), from non-genic SNPs only (method 4) or from genic SNPs in brain-expressed genes only (method 5). The effective number of SNPs denotes the number of independent SNPs that is consistent with the empirical mean and variance of the distribution of the test statistic under the null hypothesis of no association[26] (see supplementary material, section 2). Matching for the number of genes as well as for the effective number of SNPs in a functional gene group would be ideal but is highly limited as there will only be a few (< 5) sets of gene groups that can be created when the original group of genes is large (1026 in our case). We thus created matched control groups following the five methods described above, each testing slightly different null hypotheses (see Table 1).

--- insert Table 1 ---

For each of the five competitive test designs, 100 matched control groups were drawn. For each draw we carried out an association analysis of all SNPs in the matched control group in each of the three datasets, calculated the Σ-log(P) and then conducted 10,000 permutations of the dataset to determine the empirical P-value of each of the 100 matched control groups of genes in each dataset, similar to the actual analysis with the group of synaptic genes. These empirical P-values were combined across the three datasets using Stouffer's weighted Z-transform method[23]. For the five control designs, we thus obtained five sets of 100 combined empirical *P*-values. We then calculated how often the true combined empirical *P*-value (from the synaptic gene group) was higher than the combined empirical *P*-value from the matched control groups and divided that by the number of draws. Since there were 100 draws, the lowest empirical *P*-value of the combined empirical *P*-value that could be obtained was <.01, when none of the combined empirical *P*-values from the random draws was equal or more significant than the combined empirical *P*-value from the synaptic gene group (see Figure 1 for a graphical overview of the steps in data analysis).

--- insert Figure 1 ---

Enrichment tests of previously implicated genes

To test whether synaptic functional groups contained previously implicated genes more often than by chance we retrieved all SNPs with P ≤ 1.0-5 from all significant loci reported in GWA studies for schizophrenia that were published before 14th of February 2011, using the GWAS catalogue[27] and mapped these loci to protein coding genes (NCBI build v36.3). In addition we added genes implicated from genome-wide CNV studies. Fisher exact tests were used to determine the presence of enrichment.

Results

Are synaptic genes significantly associated with the risk of schizophrenia?

No individual SNP reached the threshold for genome-wide significance in any of the three datasets using a genome wide association analysis for each dataset separately (see supplementary material, Figure S1). Functional gene group analysis of all 1026 synaptic genes jointly resulted in a significant association of the total group of pre- and postsynaptic genes to the risk of schizophrenia. This was true in all three samples separately and highly significant when combined across samples, with a combined *P*-value of 7.6×10^{-11}. For each sample, the Σ-log(P) obtained from the original analysis with all synaptic genes was in the higher end of the empirical distribution and highly significant with only one of 10,000 permutations exceeding the observed Σ-log(P) for the ISC_affy5, ISC_affy6, and GAIN datasets (see Figure 2).

--- insert Figure 2 ---

Are synaptic genes more significantly associated with the risk of schizophrenia than randomly drawn groups of genes/genetic variants?

Results from the five control methods show that SNPs in synaptic genes are more strongly associated with the risk of schizophrenia than any other set of randomly drawn genes. For none of the control methods we found a combined empirical P-value that was more significant than the combined empirical P-value from the synaptic genes (see Figure 3). The 'empirical *P*-value of the combined empirical *P*-value' was <0.01 in all methods (see supplementary material, Tables S1 and S2), suggesting that the group of synaptic genes is generally more strongly associated with schizophrenia than other groups of genes that either include the same number of genes, the same

effective number of nongenic or genic SNPs, nongenic SNPs only, genic SNPs only, or the same effective number of SNPs drawn from brain expressed genes only.

Which synaptic subgroups are most strongly related to the risk of schizophrenia?

We tested seventeen synaptic subgroups and one group of synaptic genes that did not share a known function for association with schizophrenia. We found that three synaptic subgroups were significantly associated with increased risk of schizophrenia under the null hypothesis that none of the SNPs in these groups were associated with schizophrenia: *intracellular signal transduction group* (P=0.0002), genes related to *excitability* (P=0.0009) and genes involved in *CAT signaling* (P=0.0024) (see Table 2). The matched control methods for these subgroups resulted in P-values between .02 and .04 for the *intracellular signal transduction* group, P-values between <.01 and .03 for the *excitabilit*y group and between .03 and .06 for the *CAT signaling* group (see Figure 3).

--- insert Table 2 ---

--- insert Figure 3 ---

The signal of the most significant functional group (*intracellular signal transduction*) was mainly derived from the two ISC samples, whereas the GAIN sample contributed less to the overall evidence of significance of these groups but contributed mostly to association with the *CAT signaling* group. For the group of *excitability* genes however, all three samples independently showed nominally significant or suggestive evidence. Quantile-quantile (Q-Q) plots (supplementary Figures S3a-c) of the significant functional groups in the three samples show that for each functional group a multitude of SNPs in multiple genes, each of small effect, contribute to the overall significance, suggesting that the association cannot be explained by only a few genes in the group but rather by the joint effect of many genes in the functional group.

Synaptic subgroups include genes associated previously with schizophrenia

We tested whether the synaptic functional groups included genes for schizophrenia previously implicated from GWAS or CNV studies. The *intracellular signal transduction* group includes *NRGN* which was one of the most significant genes identified in the SGENE+ based GWAS[6], but was not below the genome-wide threshold in the ISC or MGS GWAS[5]. The *excitability* group contains *CACNA1C*, which was one the two most significant genes identified in a recent GWAS for bipolar disorder[28], and was recently also associated with schizophrenia[29]. From the group of genes involved in *CAT signaling*, four genes were implicated previously in schizophrenia. Enrichment analysis of previously implicated genes in schizophrenia from GWAS and CNV studies indicated significant, though moderate enrichment of previously associated genes in the total group of in synaptic genes (P=0.02) using a Fisher exact test. This enrichment was mainly due to enrichment in the *CAT signaling* group (P=0.0002). However, three out of the four genes in *CAT signaling group* that were implicated previously were very large genes. As significant results from GWAS studies may be biased towards large genes, the enrichment test for the *CAT signaling* group needs to be interpreted with caution (see supplementary materials, section 5).

Discussion

Our over-arching goal was to test whether genetic variation associated with schizophrenia risk accumulates in functional gene groups operating in the synapse. We showed that the total group of genes encoding proteins in the synapse was highly associated with the risk of developing schizophrenia with a combined P-value of $< 7.6\text{x}10^{-11}$. In addition, the group of synaptic genes was more strongly associated with schizophrenia than any of the matched-control groups of genes ($P < .01$). The functional gene group approach is a novel approach in which genes are grouped according to cellular function, and which goes across traditionally defined biological pathways, also referred to as horizontal versus vertical grouping[11]. We used a manually curated database of functional gene

groups, which tends to include more updated annotation information of gene function - especially for genes expressed in brain - than some of the online available databases. We do note, however, that gene function annotation is an ongoing endeavor and that annotation of functional gene groups is therefore continuously improved.

Apart from testing all genetic variants in synaptic genes as a group, we tested subgroups of synaptic functioning and found that three subgroups of synaptic functioning mainly drive the association of the synaptic gene group with schizophrenia; *Intracellular Signal Transduction* (*P*=0.0002), *Excitability* (*P*=0.0009) and *CAT signaling* (*P*=0.0024). In general these associations were stronger than associations with matched-control groups of genes, except for *CAT signaling*"(method 5, P =.06), indicating that at least some groups of similar size as the *CAT signaling* group and existing of SNPs in brain-expressed genes are more significantly associated with schizophrenia than the CAT signaling group. We do note however that the CAT signaling group overlaps with the CAM pathway from the KEGG database which was previously associated with schizophrenia in the ISC[12].

The group of *Intracellular Signal Transduction* was most strongly associated with the risk of schizophrenia and includes the *NRGN* gene*,* which was one of the most significant loci identified in the (independent) SGENE+ GWAS[6] , but – as a single marker effect - not below the threshold of significance in the individual samples on which the current analysis was based. In the samples included in our study, each individual SNP in the *Intracellular Signal Transduction* group contributed very little to the risk of schizophrenia. However, combining their contributions resulted in a significant association.

Intracellular signal transduction in neurons and synapses is characterized by a high degree of crosstalk. A great variety of initial steps, such as activation of many different cell membrane receptors, leads to changes in a rather limited number of enzymes that generate second messengers (adenylyl cyclase, phospholipases) and a limited number of second messengers inside the cell

(calcium, cyclic adenosine-monophosphate, cyclic guanosine monophosphate, inositol 1,4,5-triphosphate; reviewed in[30]). Hence, it is plausible that genetic variation in the genes encoding these factors has similar biological consequences and additive contributions to pathogenesis.

The second most significant functional group (*Excitability*) regulates steady-state and action potential-induced ionic currents and membrane potential. Many different channels can contribute but they all allow a limited number of types of biologically relevant ions to pass. Hence, as for the group of intracellular signal transduction genes, it is plausible that genetic variation in the genes encoding these channels have similar biological consequences in cellular excitability, and thus additive contributions to pathogenesis.

For complex traits with evidence for large numbers of variants of small effect size contributing to disease risk - such as schizophrenia[4, 31], multiple sclerosis[32], and type 1 diabetes mellitus[33] - it is of crucial importance to test whether a reported association with a group of genes is merely due to the polygenic nature of the disease or to the biological function of that group of genes. Any large group of genes is likely to emerge from pathway or functional gene group analysis merely because of background polygenic effects to the risk of disease. Reporting a significant association with the group of synaptic genes may therefore seem rather trivial, as it merely confirms synaptic genes are included in the multitude of genes related to schizophrenia. A more interesting question is thus whether the group of synaptic genes is more strongly related to schizophrenia than other, randomly drawn groups of genes. To test this we designed five methods in which we created matched-control groups of genetic variants. As the genetic variants were drawn from different pools, every control method tested slightly different null hypotheses, providing insight into how important an observed association with a group of genes is under a polygenic model of inheritance. We propose that such competitive tests for pathways or functional groups need to be included in any future pathway or functional gene group analysis.

In this study we investigated whether the accumulated effects of genetic variants in multiple genes may cause dysfunction of a biological system (e.g. intracellular signal transduction), while a single genetic variant is not sufficient to cause disease. The functional gene groups we defined are characterized by redundancy, which is most likely accomplished by previous gene duplication. Over time, genetic mutations may arise causing different or less optimal protein function, which may thrive in a gene pool, thus leading to diversity or genetic heterogeneity. To some extent genes in the same functional group may functionally replace each other when others function suboptimal. Such redundancy and heterogeneity provide for fail-safe mechanisms, which render functional gene groups - like most other biological systems - robust. Robustness is a property that allows a system to maintain its functions against internal and external perturbations[34,35].

Typically, in different individuals a different set of mutations may be responsible for dysfunction. As a consequence, individuals with the same disease may have completely different genetic backgrounds, which is consistent both with a polygenic model of disease and with a threshold model of disease but seriously hampers single-marker GWAS analysis, as it decreases the effect sizes of single SNPs/genes. When focusing on a functional gene group, it becomes less relevant which particular genes carry a mutation, while the number of genes carrying a mutation before the system starts to dysfunction is much more important. Robustness, inherent to for instance synaptic protein networks, and their underlying genes, may thus provide biologically meaningful ways to interpret the notion that 'thousands of genes underlie complex traits' and may provide important insights in the biological systems important in disease etiology (see supplementary material Figure S7).

Our current results suggest that multiple genes involved in synaptic functioning are important for schizophrenia, provide support for the synaptic hypothesis of schizophrenia[15,16,17,18], and provide tentative evidence for the involvement of the biological mechanisms involved in *intracellular signal*

transduction, *excitability* and *cell adhesion and trans-synaptic signaling molecules* in schizophrenia.

Conflicts of interest

No conflicts declared

Acknowledgements

Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003). The genotyping of the samples was provided through the Genetic Association Information Network (GAIN). The dataset(s) used for the analyses described in this manuscript were obtained from the GAIN Database found at http://view.ncbi.nlm.nih.gov/dbgap-controlled through dbGaP accession number phs000021.v2.p1. Samples and associated phenotype data for the genome-wide association of schizophrenia study were provided by the Molecular Genetics of Schizophrenia Collaboration (PI: P.V. Gejman, Evanston Northwestern Healthcare (ENH) and Northwestern University, Evanston, Illinois, USA). We gratefully acknowledge financial support of the NWO/VIDI (452-05-318 to DP), TOP ZonMW (40-00812-98-07-032 to LNC and MV), NWO-ALW Pilot grant (051.07.004 to MV), FP7 HEALTH-F2-2009-241498 (Eurospin consortium), Neuroscience Campus Amsterdam and the European Union Seventh Framework Program under grant agreement no. HEALTHF2-2009-242167 ('SynSys' project). ABS and MV are supported by the Centre for Medical Systems Biology (CMSB). PMV acknowledges funding from the Australian National Health and Medical Research Council (NHMRC grants 389892, 613672). Collaboration between DP and PMV was supported through a Visiting Professorship grant from the Royal Netherlands Academy of Arts and Sciences (KNAW).

The International Schizophrenia Consortium:

Susanne Akterin[9] Kristen Ardlie[3] M. Helena Azevedo[28] Nicholas Bass[6] Douglas H. R. Blackwood[7] Celia Carvalho[11] Kimberly Chambert[2,3] Khalid Choudhury[6] David Conti[11] Aiden Corvin[8] Nick J. Craddock[5] Caroline Crombie[21] David Curtis[20] Mark J. Daly[2,3,4] Susmita Datta[6] Emma Flordal Thelander[9] Eva Fredriksson[9] Stacey B. Gabrie[l3] Casey Gates[3] Lucy Georgieva[5] Michael Gill[8] Hugh Gurling[6] Peter A. Holmans[5] Christina M. Hultman[9,10] Ayman Fanous[11] Gillian Fraser[21] Elaine Kenny[8] George K. Kirov[5] James A. Knowles[11] Robert Krasucki[6] Joshua Korn[3,4] Soh Leh Kwan[12] Jacob Lawrence[6] Paul Lichtenstein[9] Antonio Macedo[28] Stuart Macgregor[14] Alan W. Maclean[7] Patrik Magnusson[9] Scott Mahon[3] Pat Malloy[7] Kevin A. McGhee[7] Andrew McQuillin[6] Helena Medeiros[11] Frank Middleton[23] Vihra Milanova[16] Christopher Morley[23] Derek W. Morris[8] Walter J. Muir[7] Ivan Nikolov[5] N. Norton[5] Colm T. O'Dushlaine[8] Michael C. O'Donovan[5] Michael J. Owen[5] Carlos N. Pato[11] Carlos Paz Ferreira[27] Ben Pickard[7] Jonathan Pimm[6] Shaun M. Purcell[1,2,3,4] Vinay Puri[6] Digby Quested[19¤] Douglas M. Ruderfer[1,2,3,4] Edward M. Scolnick[2,3] Pamela Sklar[1,2,3,4] David St Clair[12] Jennifer L. Stone[1,2,3,4] Patrick F. Sullivan[13] Emma F. Thelander[9] Srinivasa Thirumalai[18] Draga Toncheva[15] Margaret Van Beck[7] Peter M. Visscher[14] John L. Waddington[17] Nicholas Walker[22] H. Williams[5] Nigel M. Williams[5]

1 Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America,

2 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America,

3 Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America,

4 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America,

5 School of Medicine, Department of Psychological Medicine, School of Medicine, Cardiff

University, Cardiff, United Kingdom,

6 Molecular Psychiatry Laboratory, Department of Mental Health Sciences, University College London Medical School, Windeyer Institute of Medical Sciences, London, United Kingdom,

7 Division of Psychiatry, School of Molecular and Clinical Medicine, University of Edinburgh, Edinburgh, United Kingdom,

8 Neuropsychiatric Genetics Research Group, Department of Psychiatry and Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland,

9 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden,

10 Department of Neuroscience, Psychiatry, Ulleråker, Uppsala University, Uppsala, Sweden,

11 Center for Genomic Psychiatry, University of Southern California, Los Angeles, California, United States of America,

12 Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen, United Kingdom,

13 Departments of Genetics, Psychiatry, and Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,

14 Queensland Institute of Medical Research, Brisbane, Queensland, Australia,

15 Department of Medical Genetics, University Hospital Maichin Dom, Sofia, Bulgaria,

16 Department of Psychiatry, First Psychiatric Clinic, Alexander University Hospital, Sofia, Bulgaria,

17 Molecular and Cellular Therapeutics and RCSI Research Institute, Royal College of Surgeons in Ireland, Dublin, Ireland,

18 West Berkshire NHS Trust, Reading, United Kingdom,

19 West London Mental Health Trust, Hammersmith and Fulham Mental Health Unit and St Bernard's Hospital, London, United Kingdom,

20 Queen Mary College, University of London and East London and City Mental Health Trust,

Royal London Hospital, Whitechapel, London, United Kingdom,

21 Department of Mental Health, University of Aberdeen, Aberdeen, United Kingdom,

22 Ravenscraig Hospital, Inverkip Road, Greenock, United Kingdom,

23 State University of New York – Upstate Medical University, Syracuse, New York, United States of America,

24 Washington VA Medical Center, Washington D. C., United States of America,

25 Department of Psychiatry, Georgetown University School of Medicine, Washington D. C., United States of America,

26 Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, United States of America,

27 Department of Psychiatry, Sao Miguel, Azores, Portugal,

28 Department of Psychiatry, University of Coimbra, Coimbra, Portugal,

¤ Current address: Department of Psychiatry, University of Oxford, Warneford Hospital, Headington, Oxford, United Kingdom]

References

1. Andreasen NC. Symptoms, signs, and diagnosis of schizophrenia. *Lancet* 1995; **346:** 477-481.
2. Mowry BJ, Holmans PA, Pulver AE, Gejman PV, Riley B, Williams NM *et al.* Multicenter linkage study of schizophrenia loci on chromosome 22q. *Mol Psychiatry* 2004; **9:** 784-795.
3. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* 2003; **60:** 1187-1192.
4. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460:** 748-752.
5. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; **460:** 753-757.
6. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D *et al.* Common variants conferring risk of schizophrenia. *Nature* 2009; **460:** 744-747.
7 Kim Y, Zerwas S, Trace SE, Sullivan PF. Schizophrenia genetics: where next? *Schizophr Bull* May; **37:** 456-463.
8. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008; **92:** 265-272.
9. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* 2009; **85:** 13-24.
10. Jia P, Wang L, Meltzer HY, Zhao Z. Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizophr Res* 2010; **122:** 38-42.

11. Ruano D, Abecasis GR, Glaser B, Lips ES, Cornelisse LN, de Jong AP *et al.* Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *Am J Hum Genet* 2010; **86:** 113-125.

12. O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, Morris D *et al.* Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol Psychiatry* 2011**; 16:**286-292.

13. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 2007; **23:** 3251-3253.

14. Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet* 2007; **81:** 1278-1283.

15. Owen MJ, O'Donovan MC, Harrison PJ. Schizophrenia: a genetic disorder of the synapse? *BMJ* 2005; **330:** 158-159.

16. Harrison PJ, Weinberger DR. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol Psychiatry* 2005; **10:** 40-68; image 45.

17. Hayashi-Takagi A, Sawa A. Disturbed synaptic connectivity in schizophrenia: convergence of genetic risk factors during neurodevelopment. *Brain Res Bull* 2010; **83:** 140-146.

18. Johnson RD, Oliver PL, Davies KE. SNARE proteins and schizophrenia: linking synaptic and neurodevelopmental hypotheses. *Acta Biochim Pol* 2008; **55:** 619-628.

19. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010; **11:** 843-854.

20. Suarez BK, Duan J, Sanders AR, Hinrichs AL, Jin CH, Hou C *et al.* Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample. *Am J Hum Genet* 2006; **78:** 315-333.

21. Li K, Hornshaw MP, van Minnen J, Smalla KH, Gundelfinger ED, Smit AB. Organelle proteomics of rat synaptic proteins: correlation-profiling by isotope-coded affinity tagging in conjunction with liquid chromatography-tandem mass spectrometry to reveal post-synaptic density specific proteins. *J Proteome Res* 2005; **4:** 725-733.

22. Takamori S, Holt M, Stenius K, Lemke EA, Gronborg M, Riedel D *et al.* Molecular anatomy of a trafficking organelle. *Cell* 2006; **127:** 831-846.

23. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol* 2005; **18:** 1368-1373.

24. Fisher RA. *Statistical methods for research workers*. Oliver and Boyd: London, 1925.

25. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. *Genet Epidemiol* 2002; **22:** 170-185.

26. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74:** 765-769.

27. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009; 106: 9362-9367.

28. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* 2008; **40:** 1056-1058.

29. Green EK, Grozeva D, Jones I, Jones L, Kirov G, Caesar S *et al.* The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol Psychiatry* 2010; **15:** 1016-1022.

30. de Jong AP, Verhage M. Presynaptic signal transduction pathways that modulate synaptic transmission. *Curr Opin Neurobiol* 2009; **19:** 245-253.

31. The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-Wide Association Study Identifies Five Novel Schizophrenia Loci. *Nat Genet*, in press

32. Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, Pericak-Vance MA *et al.* Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. *Am J Hum Genet* 2010; **86:** 621-625.

33. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* 2009; **5:** e1000678.

34. Kitano H. Biological robustness. *Nat Rev Genet* 2004; **5:** 826-837.

35. Stelling J, Sauer U, Szallasi Z, Doyle FJ, 3rd, Doyle J. Robustness of cellular functions. *Cell* 2004; **118:** 675-685.

Table 1 Five applied competitive control methods to test whether synaptic genes are more strongly associated to the risk for schizophrenia than any other set of randomly grouped genes or SNPs

Method	Matched for	SNPs drawn from	Null hypothesis
1	Number of genes	Genic SNPs, excluding SNPs in synaptic genes	No more evidence for association in the group of synaptic genes than any other set of an equal number of genes
2	Effective number of SNPs	Genic and non-genic SNPs, excluding SNPs in synaptic genes	No more evidence for association in the group of synaptic genes than any other set of an equal effective number of SNPs
3	Effective number of SNPs	Genic SNPs, excluding SNPs in synaptic genes	No more evidence for association in the group of synaptic genes than any other set of an equal effective number of genic SNPs
4	Effective number of SNPs	Non-genic SNPs	No more evidence for association in the group of synaptic genes than any other set of an equal effective number of non-genic SNPs
5	Effective number of SNPs	Genic SNPs in genes expressed in brain, excluding SNPs in synaptic genes	No more evidence for association in the group of synaptic genes than any other set of an equal effective number of genic SNPs from brain-expressed genes

Table 2 Association of synaptic functional gene groups with schizophrenia in three datasets

	ISC AFFY5				ISC AFFY6				GAIN AFFY6				ALL
	N genes	N SNPs	Σ -log(P)	P_{EMP}	N genes	N SNPs	Σ -log(P)	P_{EMP}	N genes	N SNPs	Σ -log(P)	P_{EMP}	P_{COMB}
All synaptic genes	795	15105	7102	1.0E-04	906	34860	16071	1.0E-04	908	35412	16348	1.0E-04	7.6E-11
Intracellular signal transduction	112	2350	1140	0.0061	133	5387	2590	0.0037	134	5475	2450	0.2088	0.0002
Excitability	47	1120	555	0.0233	50	2656	1238	0.1026	50	2680	1327	0.01	0.0009
CAT signaling	69	3278	1483	0.1568	79	7866	3564	0.1181	79	7962	3888	0.0013	0.0024
Endocytosis	20	257	133	0.1000	23	576	312	0.0379	23	581	312	0.0384	0.0029
Structural plasticity	72	1169	547	0.1388	81	2463	1172	0.0599	82	2494	1171	0.0887	0.0108
GPCR signaling	31	786	395	0.0395	37	1823	879	0.0832	37	1844	823	0.334	0.0162
'Unknown'	45	556	266	0.1250	50	1514	786	0.0087	50	1542	641	0.7143	0.0272
Protein cluster	40	893	463	0.0160	46	2039	1001	0.0435	46	2069	840	0.821	0.0272
Tyrosine kinase signaling	7	334	185	0.0190	7	786	352	0.3401	7	798	356	0.3802	0.0491
Cell metabolism	41	216	112	0.0581	44	473	198	0.6137	44	477	224	0.183	0.1154
Neurotransmitter metabolism	25	258	113	0.4492	27	580	290	0.0832	27	606	274	0.3137	0.1166
Intracellular trafficking	59	455	205	0.3364	70	987	424	0.5422	69	1020	519	0.0286	0.1329
LGIC signaling	33	842	403	0.1210	35	1910	820	0.5437	35	1948	855	0.4308	0.2370
Exocytosis	68	1170	512	0.4321	78	2547	1084	0.6207	78	2589	1170	0.2500	0.4062
RPSFB	47	376	160	0.5549	62	869	403	0.1933	63	889	368	0.7077	0.4204
Ion balance/transport	35	322	129	0.7092	41	727	295	0.7088	41	751	372	0.1044	0.5266
Peptide/Neurothropin signals	21	515	221	0.5124	21	1122	427	0.8919	21	1132	530	0.2133	0.6615
G-protein relay	23	208	81	0.7922	23	536	235	0.4444	23	556	228	0.6758	0.7327

All P_{EMP} values are based on 10,000 permutations of the data, CAT = Cell adhesion and transsynaptic molecules, GPCR = G-protein coupled receptors, RPSFB = RNA and protein synthesis, folding and breakdown, LGIC = Ligand gated ion channel. Unknown = genes that are known to be expressed in the synapse but currently are have no known shared function with other genes.

Figure 1 Overview of steps in data-analysis. The arrows in red represent the flow for the real data whereas the blue arrows represent the flow for the control methods.

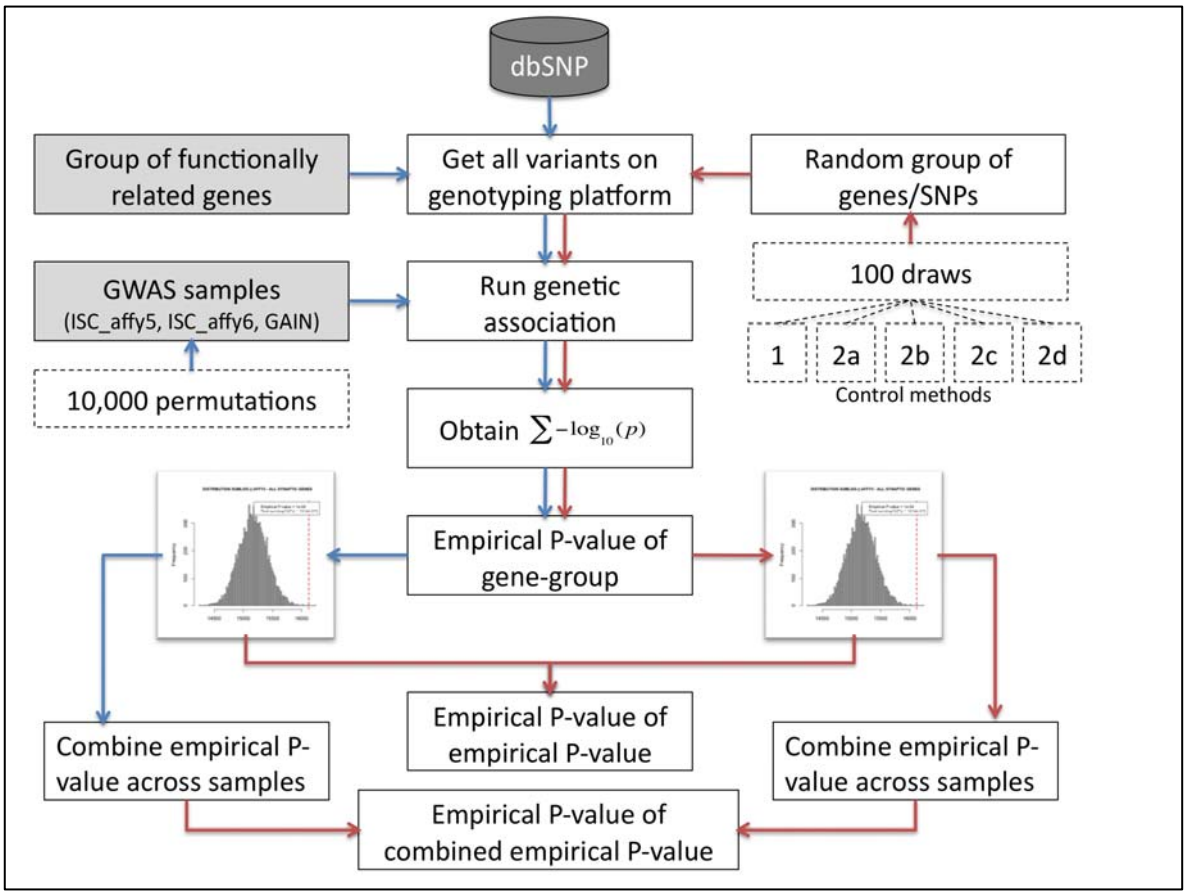


Figure 2 Empirical distribution under the null hypothesis of no association between any SNP and the risk to schizophrenia. The Σ-log(P) obtained in the original analysis of all synaptic genes is indicated in red.

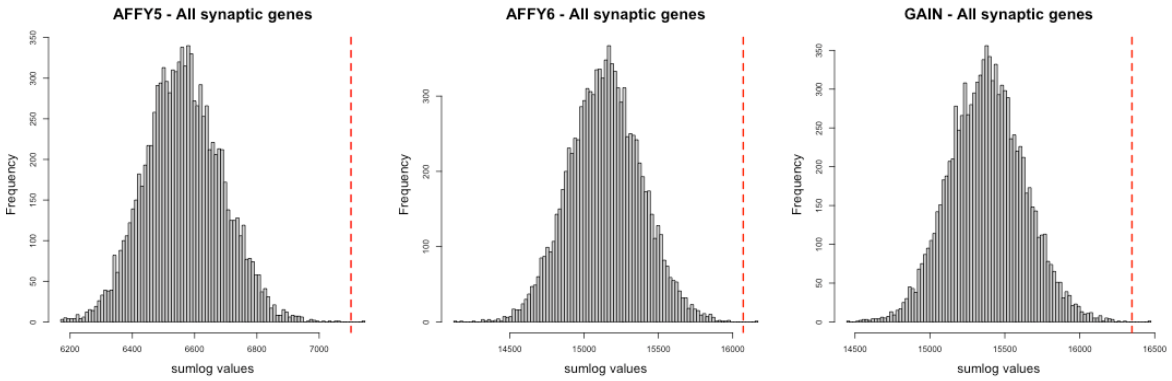
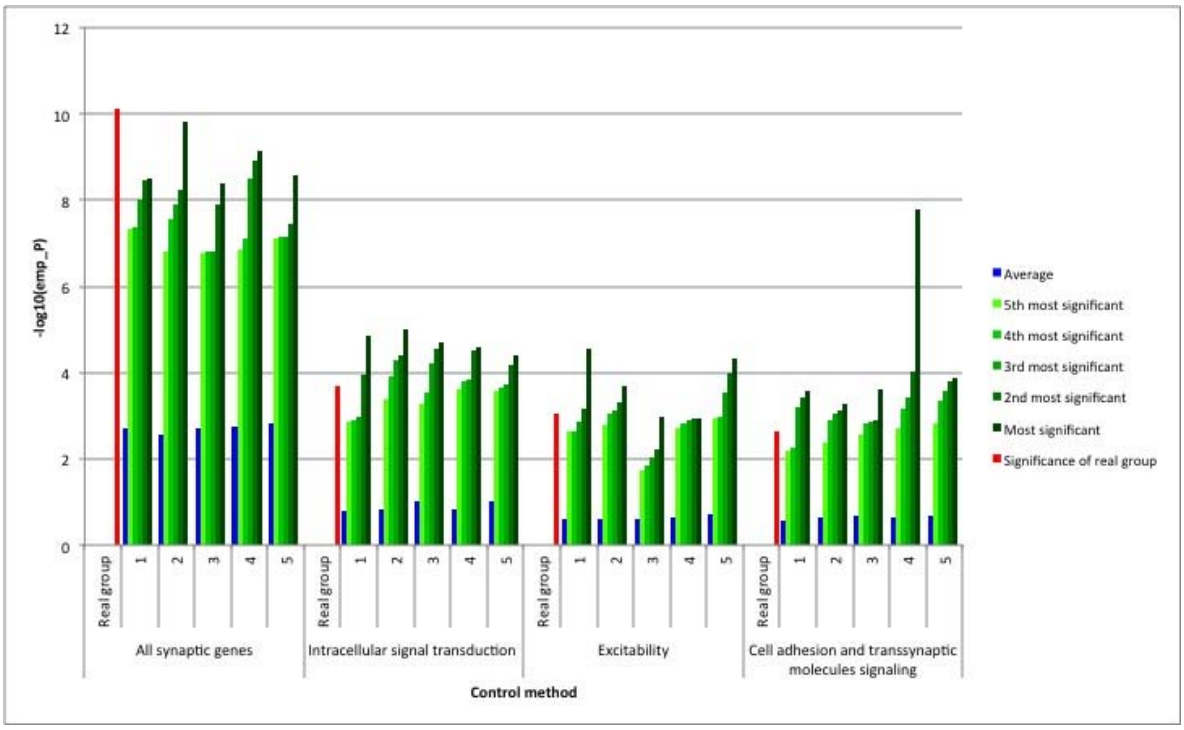
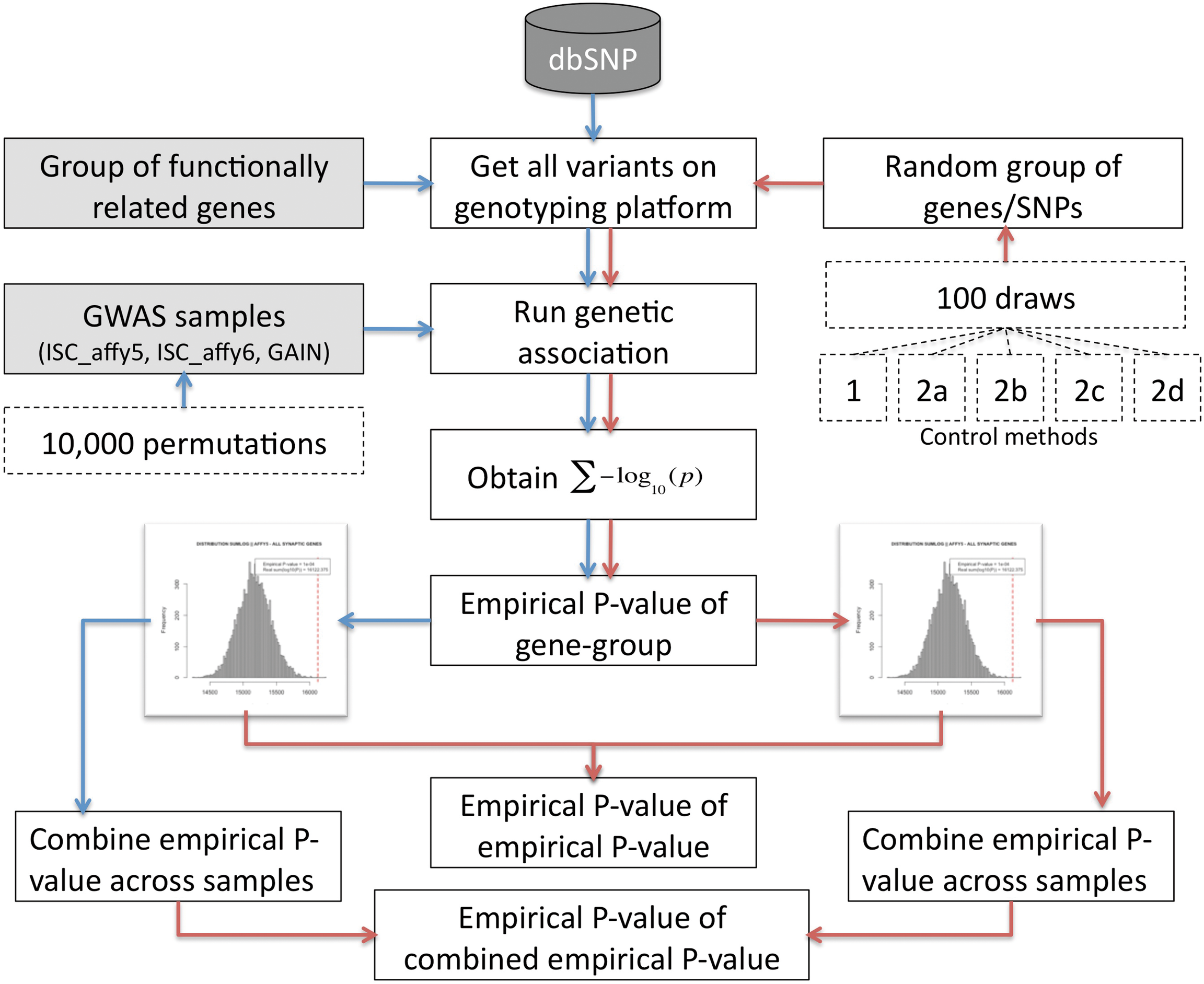
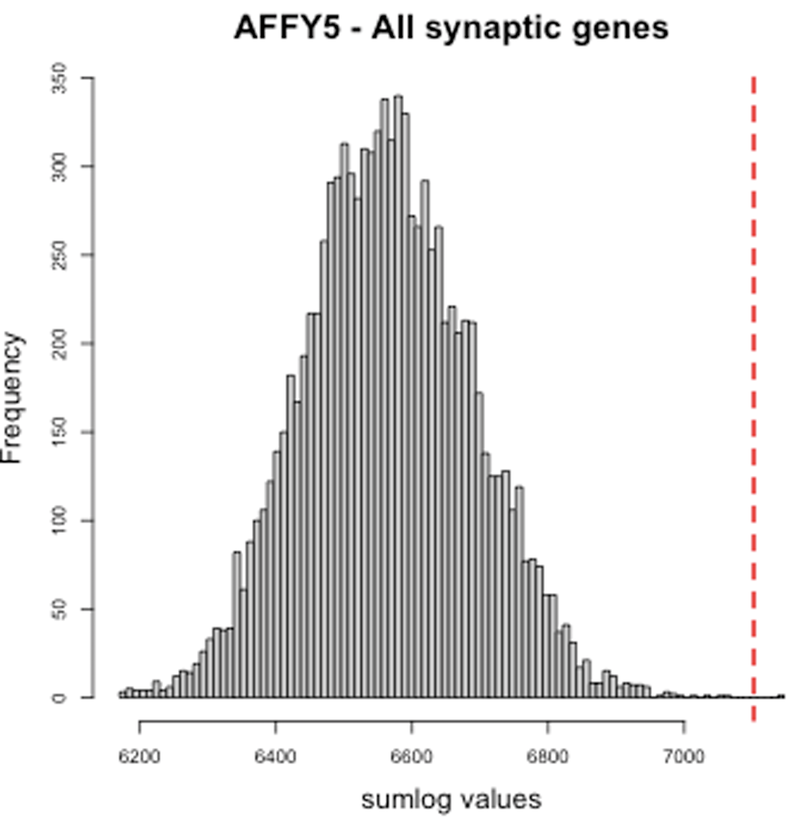


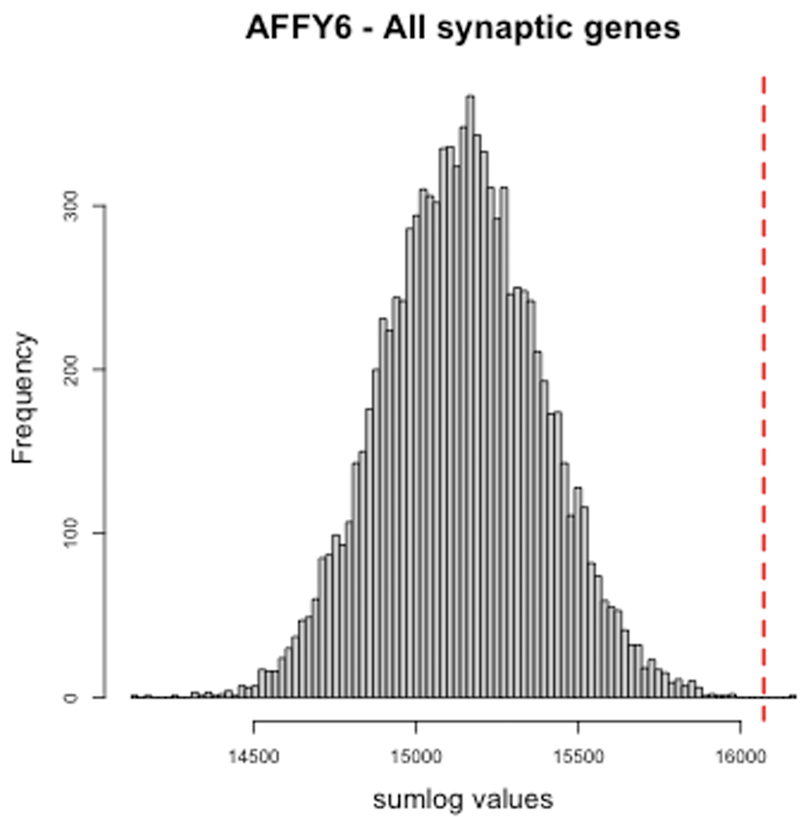
Figure 3 Overview of combined empirical *P*-values from the total group of synaptic genes and the three sub-groups that were significant after correction for multiple testing, obtained from the analysis based on the actual functional gene groups ('real', red bars), the five most significant results from the 100 draws for each control method (green bars) as well as the average combined empirical *P*-value (blue bars) obtained from five control methods across 100 draws.

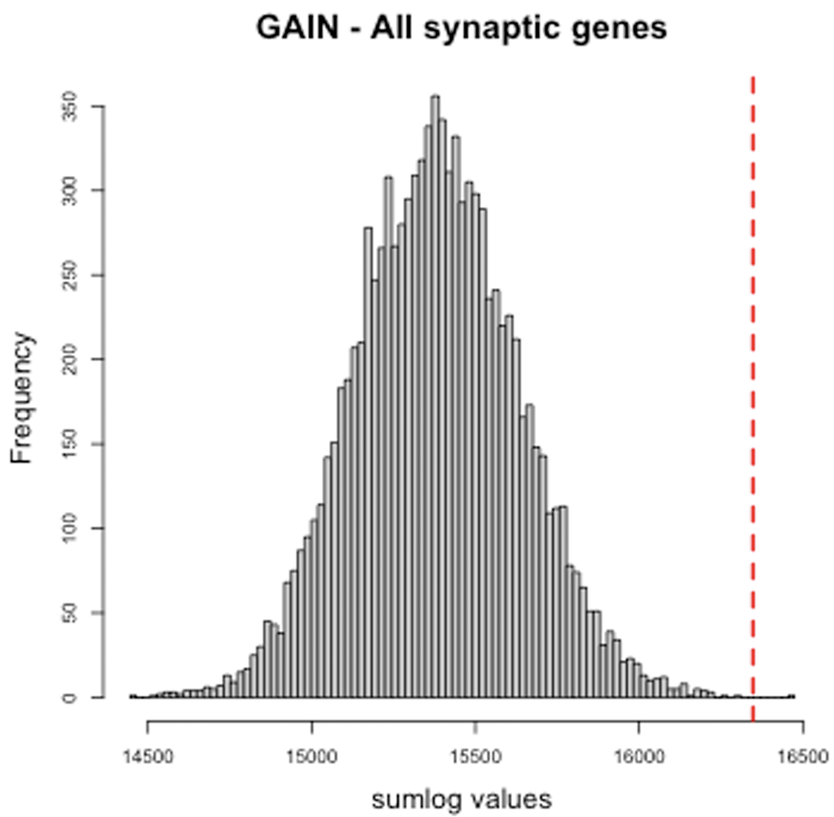


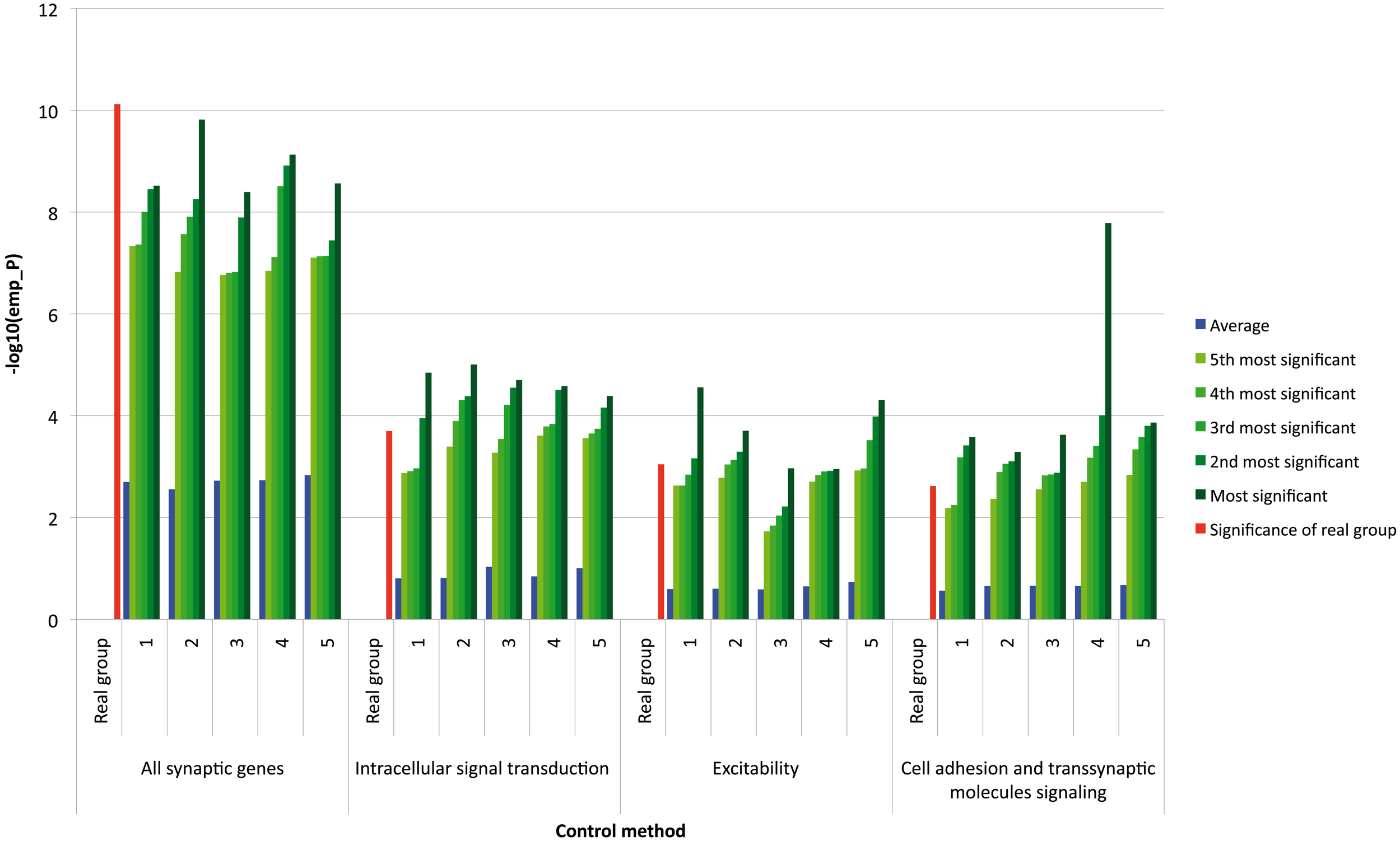
Note: The combined empirical P-values for the real group analysis as well as those for each of the 100 draws in 5 control methods, are obtained from 10 000 permutations of the data and are the combined *P*-values across the three samples. For the 'all synaptic genes' group none of the control methods resulted in a lower *P*-value than the real analysis (i.e. all empirical *P*-values of the empirical *P*-values < .01), for the intracellular signal transduction group control methods 1, 2, 3, 4, 5, resulted in empirical *P*-values of the empirical *P*-values of .02, .04, .03, .04, .03 respectively. For Excitability this was .02, .03, <.01, <.01 and .03 respectively and for cell adhesion and transsynaptic molecules signaling this was .03, .04, .04, .06., and .05 respectively. For description of competitive control methods and different null hypotheses tested we refer to Table 1.


dbSNP
Group of functionally related genes
Get all variants on genotyping platform
Random group of genes/SNPs
GWAS samples
(ISC_affy5, ISC_affy6, GAIN)
Run genetic association
100 draws
1
2a
2b
2c
2d
Control methods
10,000 permutations
Obtain $\sum -\log_{10}(p)$
Empirical P-value of gene-group
Empirical P-value of empirical P-value
Combine empirical P-value across samples
Combine empirical P-value across samples
Empirical P-value of combined empirical P-value


AFFY5 - All synaptic genes
Frequency
0
50
100
150
200
250
300
350
6200
6400
6600
6800
7000
sumlog values


AFFY6 - All synaptic genes
Frequency
0
100
200
300
14500
15000
15500
16000
sumlog values


GAIN - All synaptic genes
Frequency
0
50
100
150
200
250
300
350
14500
15000
15500
16000
16500
sumlog values


12
10
8
6
4
2
0
-log10(emp_P)
Real group
1
2
3
4
5
All synaptic genes
Intracellular signal transduction
Excitability
Cell adhesion and transsynaptic molecules signaling
Control method
Average
5th most significant
4th most significant
3rd most significant
2nd most significant
Most significant
Significance of real group