

RESEARCH ARTICLE

# Leveraging Genomic Annotations and Pleiotropic Enrichment for Improved Replication Rates in Schizophrenia GWAS

Yunpeng Wang<sup>1,2,3,4</sup>, Wesley K. Thompson<sup>5</sup>, Andrew J. Schork<sup>6</sup>, Dominic Holland<sup>4</sup>, Chi-Hua Chen<sup>4,7</sup>, Francesco Bettella<sup>1,2</sup>, Rahul S. Desikan<sup>4,7</sup>, Wen Li<sup>1,2</sup>, Aree Witoelar<sup>1,2</sup>, Verena Zuber<sup>1,2</sup>, Anna Devor<sup>3,4</sup>, Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium<sup>¶</sup>, Enhancing Neuro Imaging Genetics through Meta Analysis Consortium<sup>¶</sup>, Markus M. Nöthen<sup>8</sup>, Marcella Rietschel<sup>9</sup>, Qiang Chen<sup>10</sup>, Thomas Werge<sup>11</sup>, Sven Cichon<sup>12</sup>, Daniel R. Weinberger<sup>10</sup>, Srdjan Djurovic<sup>1,13</sup>, Michael O'Donovan<sup>14</sup>, Peter M. Visscher<sup>15,16</sup>, Ole A. Andreassen<sup>1,2\*</sup>, Anders M. Dale<sup>3,4,5,7\*</sup>



CrossMark  
click for updates

**OPEN ACCESS**

**Citation:** Wang Y, Thompson WK, Schork AJ, Holland D, Chen C-H, Bettella F, et al. (2016) Leveraging Genomic Annotations and Pleiotropic Enrichment for Improved Replication Rates in Schizophrenia GWAS. *PLoS Genet* 12(1): e1005803. doi:10.1371/journal.pgen.1005803

**Editor:** Greg Gibson, Georgia Institute of Technology, UNITED STATES

**Received:** May 14, 2015

**Accepted:** December 21, 2015

**Published:** January 25, 2016

**Copyright:** © 2016 Wang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Data from the Psychiatry Genomics Consortium for Schizophrenia and Bipolar can be accessed by following the consortium's data policies: <http://www.med.unc.edu/pgc/data-sharing>; brain structure and putamen volume data is available from the ENIGMA2 Consortium from Derrek Hibar and Paul Thompson by contacting [enigma@ini.usc.edu](mailto:enigma@ini.usc.edu); and the Crohn's disease summary statistics data are available from the International Inflammatory Bowel Disease Genetics Consortium (IBDGC) by contacting [enquiries@ibdgenetics.org](mailto:enquiries@ibdgenetics.org).

**1** NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, **2** Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway, **3** Department of Neurosciences, University of California, San Diego, La Jolla, California, United States of America, **4** Multimodal Imaging Laboratory, University of California at San Diego, La Jolla, California, United States of America, **5** Department of Psychiatry, University of California, San Diego, La Jolla, California, United States of America, **6** Department of Cognitive Sciences, University of California at San Diego, La Jolla, California, United States of America, **7** Department of Radiology, University of California, San Diego, La Jolla, California, United States of America, **8** Institute of Human Genetics, University of Bonn, Bonn, Germany, **9** Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Mannheim, Germany, **10** Lieber Institute for Brain Development, Baltimore, Maryland, United States of America, **11** Institute of Biological Psychiatry, MHC, Sct. Hans Hospital and University of Copenhagen, Copenhagen, Denmark, **12** Department of Biomedicine, University of Basel, Basel, Switzerland, **13** Department of Medical Genetics, Oslo University Hospital, Oslo, Norway, **14** MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff, United Kingdom, **15** The Queensland Brain Institute, The University of Queensland, Brisbane, Australia, **16** University of Queensland Diamantina Institute, University of Queensland, Translational Research Institute (TRI), Brisbane, Australia

¶ Memberships of the Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium and Enhancing Neuro Imaging Genetics through Meta Analysis Consortium are provided in [S1 Text](#).

\* [amdale@ucsd.edu](mailto:amdale@ucsd.edu) (AMD); [o.a.andreassen@medisin.uio.no](mailto:o.a.andreassen@medisin.uio.no) (OAA)

## Abstract

Most of the genetic architecture of schizophrenia (SCZ) has not yet been identified. Here, we apply a novel statistical algorithm called Covariate-Modulated Mixture Modeling (CM3), which incorporates auxiliary information (heterozygosity, total linkage disequilibrium, genomic annotations, pleiotropy) for each single nucleotide polymorphism (SNP) to enable more accurate estimation of replication probabilities, conditional on the observed test statistic (“z-score”) of the SNP. We use a multiple logistic regression on z-scores to combine information from auxiliary information to derive a “relative enrichment score” for each SNP. For each stratum of these relative enrichment scores, we obtain nonparametric estimates of posterior expected test statistics and replication probabilities as a function of discovery z-scores, using a resampling-based approach that repeatedly and randomly partitions meta-analysis sub-studies into training and replication samples. We fit a scale mixture of two

**Funding:** OAA was supported by the Research Council of Norway ([www.forskningsradet.no](http://www.forskningsradet.no), #213837 and #223273), the South East Norway Health Authority ([www.helse-sorost.no](http://www.helse-sorost.no), #2013-123), the Kristian Gerhard Jebsen Foundation (<http://www.stiftkgj.no/?lang=en>); AMD was supported by the National Institute of Health (<http://www.nih.gov/>, GM104400-01A). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

Gaussians model to each stratum, obtaining parameter estimates that minimize the sum of squared differences of the scale-mixture model with the stratified nonparametric estimates. We apply this approach to the recent genome-wide association study (GWAS) of SCZ ( $n = 82,315$ ), obtaining a good fit between the model-based and observed effect sizes and replication probabilities. We observed that SNPs with low enrichment scores replicate with a lower probability than SNPs with high enrichment scores even when both they are genome-wide significant ( $p < 5 \times 10^{-8}$ ). There were 693 and 219 independent loci with model-based replication rates  $\geq 80\%$  and  $\geq 90\%$ , respectively. Compared to analyses not incorporating relative enrichment scores, CM3 increased out-of-sample yield for SNPs that replicate at a given rate. This demonstrates that replication probabilities can be more accurately estimated using prior enrichment information with CM3.

## Author Summary

Genome-wide association studies (GWAS) have thus far identified only a small fraction of the heritability of common complex disorders, such as schizophrenia. Here, we demonstrate that by using auxiliary information we can improve estimates of replication probabilities from GWAS summary statistics. The proposed *Covariate-Modulated Mixture Model* (CM3) incorporates auxiliary information to construct an “enrichment score” for each single nucleotide polymorphism (SNP). We show that a scale mixture of two Gaussians provides a good fit to the observed effect size distribution stratified by the predicted enrichment score when applied the method to a recent genome-wide association study (GWAS) of SCZ ( $n = 82,315$ ). Compared to estimates performed not using auxiliary information, the CM3 more accurately models the observed replication rates by stratifying on covariate-modulated enrichment scores. We observed that SNPs with low enrichment scores replicate with a lower probability compared to SNPs with high enrichment scores, even when both are genome-wide significant ( $p < 5 \times 10^{-8}$ ). At model-based replication rates  $\geq 80\%$  and  $\geq 90\%$  there were 693 and 219 independent loci, respectively. Increased out-of-sample yield for SNPs ranked according to CM3 demonstrate the utility of incorporating auxiliary information via CM3.

## Introduction

Schizophrenia (SCZ) is one of the most heritable of human diseases, with estimates of the proportion of disease risk due to genetic factors ranging from 0.6 to 0.8[1]. However, until recently, GWAS have identified only a small number of associated genes or loci, accounting for a miniscule fraction of the heritability[2]. The turning point has been the establishment of the Psychiatric Genomic Consortium (PGC)[3], which has enabled the pooling of large numbers of independent studies, thus greatly increasing the power for identification of genes affecting disease risk, and confirming the polygenic nature of schizophrenia and other psychiatric disorders[2].

In most highly polygenic traits and diseases, individual genetic loci account for a very small portion of the phenotypic variance[4]. While increasing GWAS sample sizes is crucial, another key to improving estimates of which loci will replicate in independent studies is the application of statistical methods that incorporate auxiliary information. We have previously shown, using

GWAS summary statistics from a smaller SCZ study ( $n = 21,856$ ) [2], that genomic annotation categories [5, 6] and association with bipolar disorder (BIP) [7] significantly enriches test statistics for non-null associations. Pleiotropic enrichment was also observed between SCZ and other psychiatric and somatic phenotypes [8, 9]. Together with the ENCODE findings [10], these results provide a strong evidence against *a priori* equivalence, or statistical exchangeability, of all SNPs. These results instead suggest that the probability of association should be allowed to vary as a function of the relative enrichment of different SNP categories.

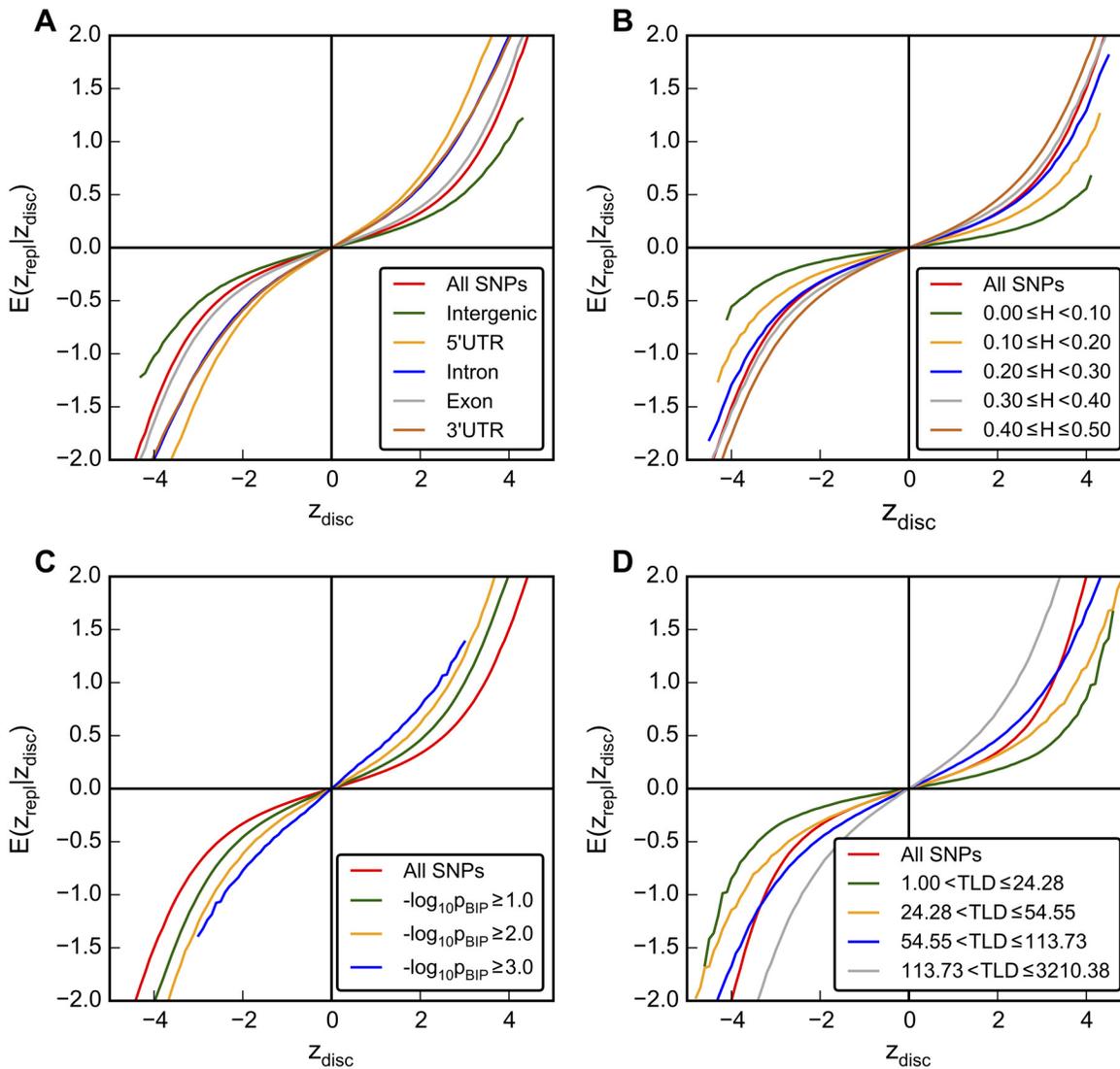
Here we present a novel algorithm, termed *Covariate Modulated Mixture Modeling* (CM3) that combines multiple sources of enrichment information to estimate SNP posterior effect sizes and to rank genetic loci based on covariate-modulated strength of association with a given trait or disease, i.e., loci that have the highest model-based estimates of probability of replication. The proposed method models thresholded z-scores as a function of enrichment categories, via logistic regression, to estimate a relative enrichment score for each SNP. Enrichment scores are then stratified into K bins: for a given enrichment stratum, we fit a scale-mixture of two Gaussians model to summary statistics within the stratum. These stratified mixture models allow for estimation of the expected z-scores and replication rates, given the observed z-scores and the effective sample size of the discovery and replication datasets. We hypothesized that sorting SNPs by the predicted replication probability from CM3 would improve out-of-sample yield, for a given replication rate, relative to the standard approach of sorting SNPs by nominal p-values alone. Here we compute the empirical replication rate as the proportion of SNPs having p values  $\leq 0.05$  in an independent sample within a set of SNPs.

We applied the CM3 method to the latest PGC SCZ sample, including  $n = 35,476$  patients with SCZ and  $n = 46,839$  controls, across 52 separate sub-studies [11]. We incorporated the following four auxiliary information categories: 1) linkage disequilibrium (LD)-weighted genomic annotations; 2) total LD (TLD); 3) heterozygosity (H); and 4) pleiotropy with bipolar disorder (BIP) (see [Materials and Methods](#)). Our results show that a stratified scale-mixture of two Gaussians model appears to fit the SCZ data well across different enrichment strata. After incorporating auxiliary information via CM3, more SNPs replicate at a given rate (e.g., 90%) compared to sorting SNPs by nominal p-values alone. Thus, enrichment methods such as CM3 may provide effective criteria for ranking SNPs in GWAS for further investigation, with potential implications for improved gene discovery and polygenic risk prediction for personalized medicine.

## Results

### Sources of Differential Enrichment

We show the results from 500 iterations of the resampling algorithm using split-halves (50% of PGC SCZ sub-studies as discovery and the other 50% as replication samples), with inverse-variance weighted meta-analysis z-scores computed for both “discovery” and “replication” samples [12]. [Fig 1A](#) shows the mean z-score across replications as a function of z-scores in the discovery samples, for different LD-weighted genomic annotation categories. For a given z-score in the discovery sample, tag SNPs in LD with enriched categories such as 5'UTR, exon, and 3'UTR variants have a higher mean z-scores in the replication sample compared to less enriched categories (e.g., intergenic; see [S2 Fig](#) for all categories studied). We also investigated the relative “enrichment” due to heterozygosity (H). [Fig 1B](#) shows mean replication z-scores as a function of z-scores in the discovery sample, for different ranges of H. For a given z-score in the discovery sample, tag SNPs with higher H have a higher mean z-score in the replication sample compared to SNPs with lower H. In addition, we calculated the mean replication sample z-scores as a function of z-scores in the discovery sample for different levels of association



**Fig 1. Mean replication z-scores stratified by genomic annotation, pleiotropy and heterozygosity.** The conditional mean z-scores in replication sample (y axis) were plotted against the z-scores in the discovery sample (x axis). The shrinkage of replication z-score is differentiated by A.) genomic annotation categories (All SNPs; Intergenic; 5' untranslated region, 5' UTR; Intron; Exon; and 3' untranslated region, 3' UTR), B.) by heterozygosity (H) intervals, C.) by associations with bipolar disorder (BIP; All SNPs;  $-\log_{10} p_{BIP} \geq 1.0$ ;  $-\log_{10} p \geq 2.0$ ; and  $-\log_{10} p \geq 3.0$ ) and D.) by total LD (TLD) intervals. All plots were generated by randomly assigning 26 of the PGC Schizophrenia sub-studies as discovery sample and 26 as replication sample (split half). The average value over 500 iterations is shown.

doi:10.1371/journal.pgen.1005803.g001

with BIP, after removing overlapping samples from the PGC BIP data[7]. For a given z-score in the discovery sample, SNPs with more significant association with BIP have a higher mean z-score in the replication sample compared to SNPs with less significant association with BIP (less enriched, Fig 1C). We also observed that the mean replication sample z-scores increases for a given discovery sample z-score as the total LD increases (Fig 1D). Taken together, this shows that, after conditioning on auxiliary information, SNP z-scores are not exchangeable in terms of association with SCZ. The properties of each SNP—LD with annotation categories (Fig 1A and S2 Fig), heterozygosity (H, Fig 1B and S7 Fig), association level with other traits (Fig 1C) and total LD (TLD, Fig 1D)—have implications regarding replicable associations with SCZ.

## Combined Differential Enrichment Score

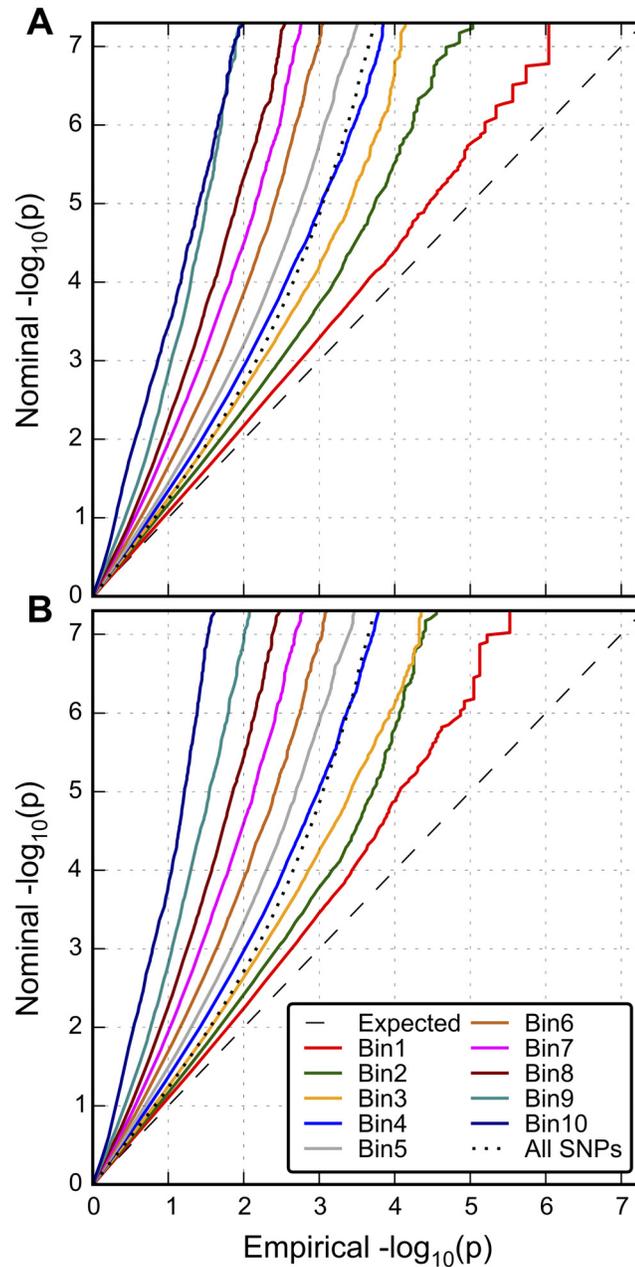
We combined information from the different sources of auxiliary information (LD-weighted annotation categories, TLD, BIP, and H) using a multiple logistic regression model, to compute the predicted relative enrichment score for each SNP (see [Materials and Methods](#)). The relative enrichment scores of all SNPs were stratified into ten equally spaced disjoint intervals. [Fig 2](#) shows the conditional Q-Q plots displaying the distribution of summary statistics for the PGC SCZ conditional on different levels of enrichment, from the least enriched stratum (Bin 1) to the most enriched stratum (Bin 10). Q-Q curves are thresholded at  $-\log_{10} p \leq 7.3$  to focus on SNPs below genome-wide significance. Comparing [Fig 2A](#) with [Fig 2B](#) shows that by including pleiotropy with Bipolar Disorder (BIP) as extra source of auxiliary information the level of enrichment increases. [Fig 3A](#) shows the stratified discovery and replication observed mean z-scores of these strata (solid lines) with split half samples, using the resampling-based strategy (see [Materials and Methods](#)). (The results for other re-sampling proportions are shown in [S3 Fig](#)). The shape of the posterior z-score functions ([Fig 3A](#)), monotonically increasing but with a relatively flat region in the middle, is characteristic of mixture distributions, with non-linear “shrinkage” towards zero (see Efron [13]). For a given z-score in the “discovery” sample, the z-scores in the “replication” sample increases with increasing degree of predicted enrichment. For example, in the case of a SNP with a z-score in the discovery sample of 2, the expected z-score in the replication sample is approximately 0.10 for the least enriched category, and approximately 1.30 for SNPs in the most enriched strata.

The solid lines in [Fig 3B](#) show the mean observed replication probabilities across random split-half partitions as a function of nominal p-values in the discovery samples, for different enrichment strata (see [Materials and Methods](#)). Results for other training/replication partition proportions are shown in [S4 Fig](#). As expected, [Fig 3B](#) shows an increase in observed replication probability with increased relative enrichment factor levels for a given p-value. For examples, for SNPs with a p-value of 0.001 ( $-\log_{10}(p) = 3.0$ ) in the discovery sample, the observed replication rate is close to 0.09 in the least enriched stratum, and increased to an observed replication rate of 0.68 for the most enriched stratum. Related to this, for a given observed replication rate, the p-value varies dramatically across enrichment strata. The corresponding Figs illustrating the observed relationships of  $z^2$  between discovery and replication samples are shown in [S5 Fig](#).

## Modeling Test Statistics and Replication Rate

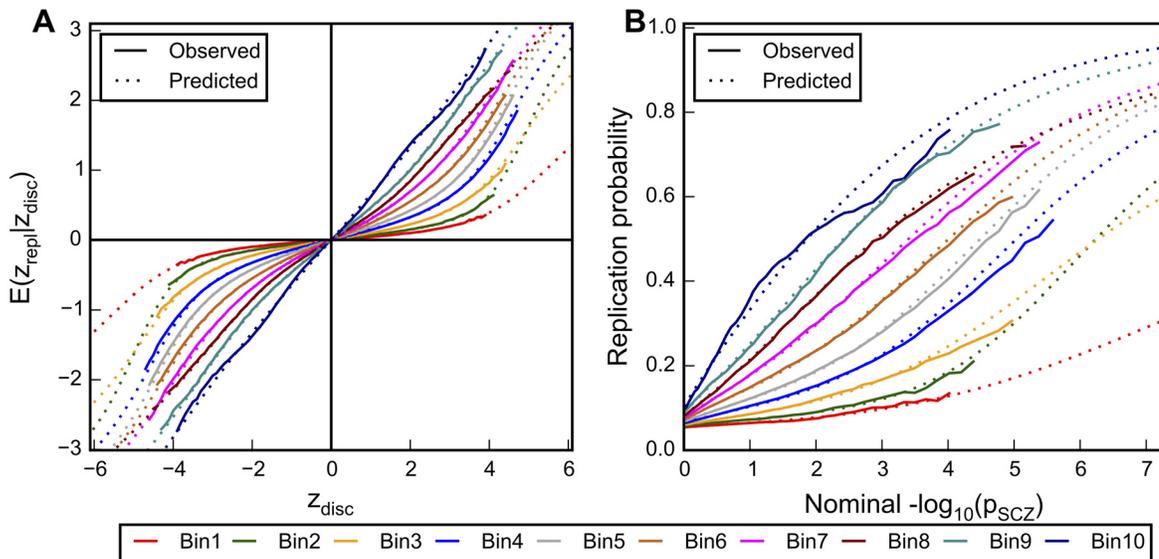
To investigate if we can model the nonparametric estimates of replication test statistic means and variances (solid lines in [Fig 3A and 3B](#)), we fit a scale-mixture of two Gaussians model to each enrichment stratum (see [Materials and Methods](#) for details). The dotted lines in [Fig 3A](#) indicate the posterior mean z-scores in replication sample as function of z-scores in discovery sample for the different enrichment strata. The corresponding observed and predicted replication probability plots are shown in [Fig 3B](#). Note that for SNPs satisfying the standard GWAS significance threshold ( $p < 5 \times 10^{-8}$ ), the predicted replication rate ranges from close to 0.28 for the least enriched stratum, to 0.94 in the most enriched stratum. In other words, SNPs obtaining the commonly used p-value threshold in GWAS ( $p = 5 \times 10^{-8}$ ) replicate more frequently if associated with a high relative enrichment score compared to SNPs with the same p-values but having a low enrichment score. The proposed mixture model appears to provide a good fit to the observed data across different enrichment levels and discovery and replication sample sizes. [S3–S5 Figs](#) show the model performance of other discovery/replication partition proportions.

To investigate the effect of sorting SNPs based on predicted replication probability instead of by nominal p-value, we computed the cumulative empirical replication rate using the



**Fig 2. Enrichment of SNP associations with schizophrenia conditioned on predicted enrichment scores.** The conditional Q-Q plot shows the enrichment of SNP association with schizophrenia stratified by predicted relative enrichment scores A.) based on LD-weighted Annotation categories, heterozygosity and total LD score and B.) based on LD-weight Annotation categories, heterozygosity, total LD score and SNP association with bipolar disorder. The predicted enrichment scores are equally divided into 10 disjoint intervals or bins (from the least enriched stratum, Bin1, to the most enriched stratum, Bin10). The dashed line indicates the null distribution and dotted line indicates all SNPs, i.e., not stratified. Different colors indicate different intervals of predicted enrichment scores. The leftward shift of the each curve compared to the null line indicates the relative enrichment. SNPs in the MHC region were excluded and then pruned based on the LD structure from the 1000 Genomes European subpopulation at  $r^2 < 0.8$ .

doi:10.1371/journal.pgen.1005803.g002

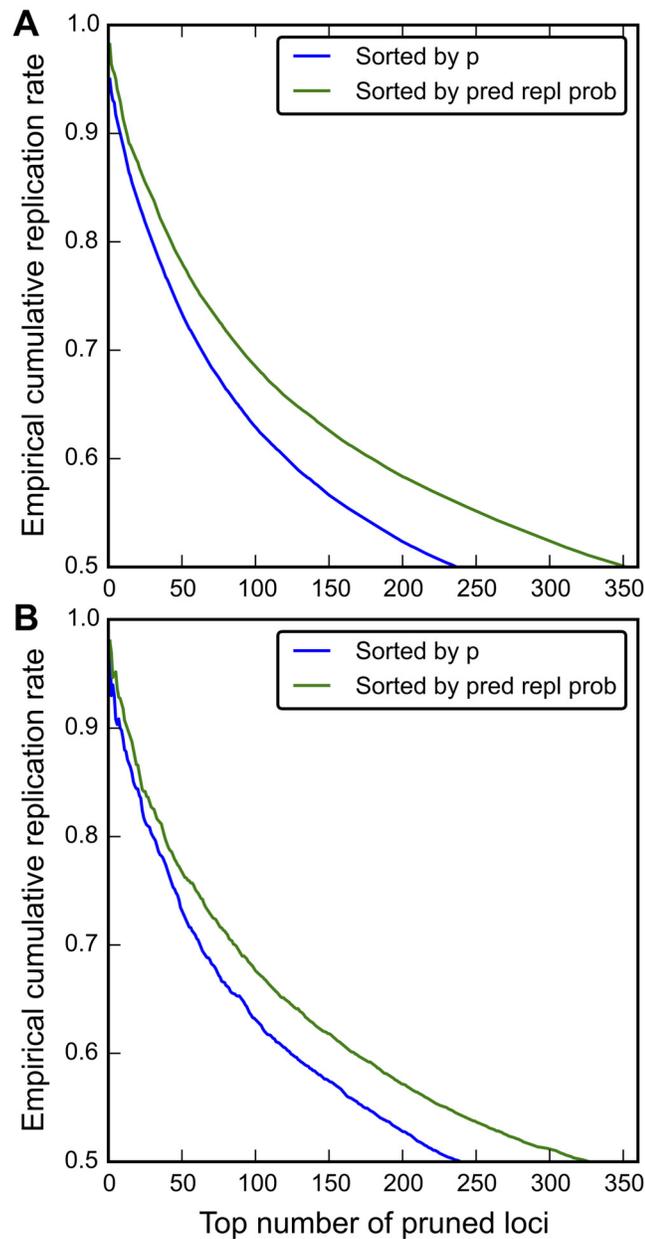


**Fig 3. Mean replication z-score and replication rate stratified by enrichment scores.** A.) The observed (solid lines) and predicted (dotted lines) mean z-scores in the replication sample (y axis) were plotted against the z-scores in the discovery sample (x axis). The shrinkage of replication z-scores is differentiated by disjoint intervals of relative enrichment scores. B.) The observed (solid lines) and predicted (dotted lines) replication probabilities were plotted against the negative common logarithm of nominal p values of schizophrenia SNPs in discovery sample (x axis). Colors indicate the 10 disjoint intervals of relative enrichment scores, ranging from the least enriched (Bin1) to the most enriched (Bin10). All data were generated by randomly assigning 26 of the PGC schizophrenia sub-studies as discovery sample and 26 as replication sample (split half). The averaged value over 500 iterations was shown.

doi:10.1371/journal.pgen.1005803.g003

random partition approach (split-half). Fig 4 shows the comparison of the observed cumulative replication rate with SNPs sorted by the predicted replication probability from CM3 (from high to low) and nominal p-values (from low to high). Using the CM3 method, a larger number of loci are selected for a given cumulative observed replication rate than when ranking SNPs by nominal p-values. For instance, for the CM3 method an average of 353 loci replicate at a replication rate of 0.5, whereas an average of 238 replicate at the same rate when using nominal p-values without auxiliary information (Fig 4A). Further, when sorted by p-values, no SNPs replicate at a rate higher than 0.95, whereas when sorted by predicted replication probability, the highest-ranked SNPs replicate at a rate of 0.982 (Fig 4A). Fig 4B shows out-of-sample performance of CM3, i.e., only the split half discovery sample was used to fit model parameters, compared to the nominal p-values based method. At a replication rate 0.5, 328 loci replicated sorted by predicted replication probability from the CM3 methods. Taken together, this shows that incorporating auxiliary information via CM3 can provide a larger yield of SNPs for a given observed replication rate.

To assign posterior effect size estimates and predicted replication probability to each SNP for the whole PGC SCZ sample, we computed a fine grid “lookup table” as a function of the observed z-scores in the discovery sample and the enrichment score (see [Materials and Methods](#)). S8 Fig compares sorting of SNPs based on the predicted replication probability vs. on nominal p-value. The change due to sorting by CM3 is most pronounced for SNPs having smaller effect sizes, i.e., larger p-values (upper right corner of S8 Fig) and less so for SNPs having smaller p-values (lower left corner of S8 Fig). We found 693 independent non-MHC loci ( $LD r^2 < 0.1$ , clumped by distance 250kbp) having predicted replication probability  $\geq 0.8$ , and 219 having predicted replication rate  $\geq 0.9$  (S1 Table). The predicted replication rate corresponding to the GWAS p-value threshold of  $5 \times 10^{-8}$  was 0.8571 in the split-half discovery/replication analysis, without stratification by relative enrichment scores (S9 Fig). At this estimated



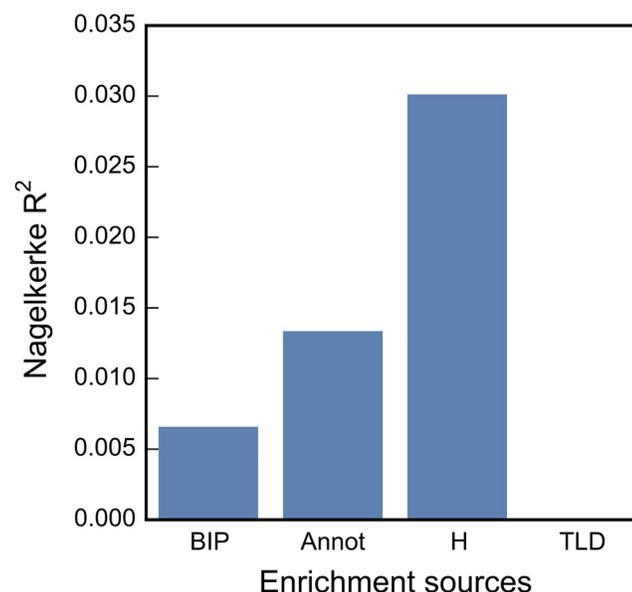
**Fig 4. CM3 improves power of identifying gene loci.** The average empirical cumulative replication rates (y axis) are plotted against the number of SNPs replicating at that rate > 0.5 (x axis), after removing MHC region SNPs and pruning at LD  $r^2 < 0.1$ . A.) The full PGC sample was used to estimate predicted replication probability (pred repl prob). For each iteration, 26 PGC schizophrenia sub-studies were randomly assigned to the discovery sample, and the rest to the replication sample (split half). The average values over 500 iterations are shown, and B.) Half of the PGC sample (26 sub-studies) was used to estimate the predicted replication probability. For each iteration, 26 PGC schizophrenia sub-studies were randomly assigned to the discovery sample, and the rest to the replication sample (split half). Then, the predicted replication probability was estimated by applying the CM3 method on the discovery sample with 50 iterations. The p values (computed by meta-analysis) of the discovery sample and the predicted replication probabilities (computed by CM3) were used to sort SNPs in replication sample, consist of rest of the sub-studies. The average replication rates across 50 iterations were shown. Colors indicate different sorting criteria (green: sorted by prediction replication probability and blue sorted by nominal p values).

doi:10.1371/journal.pgen.1005803.g004

replication threshold, CM3 identified 9 more regions than p-value based method (Fig 4). CM3 performs better when the size of discovery sample is smaller than replication sample, given that overall sample size is fixed (S6 Fig). We also repeated the analysis without including BIP as enrichment sources. The number of clumped independent non-MHC loci with predicted replication probability > 0.8 (0.9) becomes 428 and 201. We display the predicted replication probability in the sixth column of S1 Table.

### Relative Importance of Auxiliary Information Categories Using Logistic Regression

We next evaluated the relative importance of the different categories of auxiliary information using the same thresholded logistic regression framework that we employed for constructing relative enrichment strata. The models without LD-weighted annotation scores (Annot), TLD, H, and BIP were each in turn compared with the full model, i.e., including all four categories (Materials and Methods). Fig 5 shows the relative importance of each source measured by change in Nagelkerke's  $R^2$  (see S15 Fig, measured by the area under the receiver operating characteristic curve (AUC)). Annot, H, and BIP make major contributions to the enrichment of SNP association with SCZ (all having  $p < 10^{-16}$ , likelihood ratio test). We find that the contribution of TLD is reduced when including other sources of differential enrichment. We also observe the same qualitative results when varying the *p*thresh used in dichotomizing the nominal p-values (S10A Fig). The contribution of TLD increases when we instead regress on the unthresholded  $z^2$  on enrichment sources and using change in adjusted  $R^2$  (S10B Fig), though it still remains smaller than the change in variance from excluding Annot or H categories, and is similar in size to the change in  $R^2$  excluding BIP. Of note, the proportion of explained variance in the unthresholded  $z^2$  regression is much smaller compared to that of the thresholded logistic regression (S10B Fig).



**Fig 5. Relative importance of sources for enrichment.** The relative importance of different sources of enrichment (x axis) for explaining SNP association with schizophrenia was measured by the Nagelkerke's  $R^2$ . The enrichment sources were: total linkage disequilibrium (TLD); the squared z-scores of SNP association with bipolar disorder (BIP); the LD weighted genomic annotation scores (Annot); and the heterozygosity (H).

doi:10.1371/journal.pgen.1005803.g005

## Effect of Genetic Architecture Application to Other Phenotypes

To investigate the effect of different genetic architectures on the performance of CM3 we also analyzed the data for the brain structure Putamen volume[14] (26 sub-studies,  $N = 12,596$ ) and Crohn's diseases[15] (8 sub-studies, effective sample size  $N = 10,050$ ). We identified 11 independent regions with predicted replication probability  $> 0.8$  (S2 Table). Similarly, the number for Crohn's disease is 81 (S3 Table). S12–S14 Figs show the performance of CM3 on these two additional datasets.

## Discussion

We have presented a novel algorithm, called CM3, which provides more accurate estimates of predicted replication probabilities for each SNP in a GWAS. Sorting SNPs based on predicted finite-sample replication probabilities incorporating auxiliary information, rather than by nominal p-values, yields a larger number of SNPs for a given replication threshold. The improved performance was demonstrated by comparing the average number of independent loci for a given observed cumulative replication rate in a split-half random partitioning analysis (Fig 4). Sorting SNPs based on the predicted replication probability was found to dramatically increase the yield of GWAS consistently across observed cumulative replication rates, relative to nominal p-values alone. We observed a broad range of replication probabilities across enrichment strata, for a given nominal p-value. Taken together, these results further demonstrate that “all SNPs are not created equal” [5], and that by leveraging differential enrichment across SNPs, it may be possible to improve on standard GWAS methods.

We have previously shown that LD-weighted genomic annotations and pleiotropy can be used to enrich the association of SNPs with SCZ in GWAS [5, 8, 9, 16], and we here demonstrate additional increase in power due to heterozygosity. Further enrichment was observed by adding association with another psychiatric phenotype, bipolar disorder, in line with previous findings of overlapping gene loci (pleiotropy)[8]. SNPs with higher heterozygosity replicate at higher rates, which is in accordance with the concept that common variants explain a large portion of the variance in complex human phenotypes[17] (S7 Fig). We used a multiple logistic regression model to investigate the relative contribution to the association with SCZ by the different sources of auxiliary information. By accounting for the overlap of contributions from different sources in this model, we found that total LD, LD-weighted genomic annotations, pleiotropy and heterozygosity contribute substantially to the association. We find that the contribution of total LD is substantially reduced when including these other sources of differential enrichment. The current analyses suggest that total LD partially acts as a proxy for other predictors of differential power and enrichment for non-null effects.

The current work is motivated by a previous paper from our group on LD-weighted enrichment annotation factors that appear to be related to very many complex traits and diseases [5]. The CM3 algorithm is based on a novel random partitioning approach to non-parametrically estimate replication effect size means and variances of GWAS summary statistics [18], along with a scale-mixture of two Gaussians modeling framework, similar to that proposed by Zhou *et al.*[19]. Unlike other partition-based approaches [20], the stratified CM3 model enables prediction of the posterior effect size and replication probability for each SNP, incorporating relative enrichment scores. Scale mixture models have been widely employed in genetic analyses, such as in animal genetic, GWAS, and QTL analysis[21]. The model used here differs from others in that: 1) it uses only summary statistics from GWAS sub-studies; 2) the estimation algorithm fits the model to the mean replication effect sizes and variances[18], allowing for estimation of the effects of changing effective sample sizes on the model fit; 3) the incorporation of enrichment scores obtained from SNP-level auxiliary information.

Several other methods which incorporating enrichment factors into GWAS have lately been developed [22–25]. The CM3 method differs from these previous methods in that: 1) none of the previously published methods directly model the proportion of null (the small-effect component) vs. non-null (large-effect component) effect sizes as a function of annotations; 2) none of the previously reported methods directly model the apparent inflation of the null distribution, which we do through allowing the null component to consist of very small replicating effects; 3) unlike previous methods, the current methodology produces estimates of replication effect sizes and replication probabilities from the random partitioning algorithm; 4) finally, we use the empirical predicted replication probabilities of SNPs as evidence of association since we can directly compute the empirical replication rate from the resampling experiment and directly compare this with the prediction performance of the corresponding CM3 covariate-stratified mixture of scale normal models.

It is of note that over-fitting could be a serious issue if the logistic regression was fitted selecting from a large number of annotation factors, especially trait/disease specific enrichment factors. This could be prevented by applying model selection procedures that guard against including too many predictors in the regression model. It is also possible that over-fitting can be an issue due to computing relative enrichment scores from a logistic regression model fitted using the entire dataset, though we validated the results by increased out-of-sample yield in SNPs for a given predicted replication rate (Fig 4B). Choice of number of strata to use for enrichment analyses and constructing smoothed “lookup tables” from the resulting fits is also an issue, which could potentially also be addressed by measures of model fit vs. model complexity trade-off. Note, though including auxiliary information (such as pleiotropic enrichment) in the CM3 will change the relative ranking of SNPs, the replication effect size estimates will not be affected negatively, e.g., SNPs related to BIP may have larger effect size estimates but SNPs unrelated to BIP will have unchanged effect size estimates.

The proposed method requires the summary statistics from multiple independent studies. The performance not only depends on the overall sample sizes and on the sizes of discovery and replication samples but also on the genetic architecture of the complex diseases/traits. In general, it performs better when the discovery sample is smaller comparing to the replication sample given the overall sample size is fixed (S6 and S14 Figs). This also suggests that CM3 may be more powerful for under-power studies. The results of applying CM3 to the data for the brain structure Putamen volume (putamen) and Crohn’s disease (CD) show that even with 8 independent sub-studies (CD) the improvement can be large. However, with 26 sub-studies available for Putamen volume the improvement is minimal (S12–S14 Figs).

An important utility of the CM3 method may be selection of a greater proportion of relevant SNPs for gene set enrichment and biological pathway analyses, which often use a less stringent p-value threshold than the established GWAS standard for discovery. Further, the proposed method may improve the efficiency of the two-stage GWAS meta-analysis by better predicting which regions will reach significance in the combined sample, relative to the standard method of picking all SNPs that reach a significance of  $p < 1 \times 10^{-6}$  in Stage I [26, 27].

In conclusion, we have presented a novel statistical method, the covariate modulated mixture model (CM3), which incorporates multiple sources of auxiliary information, such as total LD, heterozygosity, genomic annotations, and pleiotropy, for estimating effect sizes and predicting replication rates for SNPs in independent samples. The CM3 method first creates enrichment strata via multiple logistic regression, subsequently implementing a novel resampling-based algorithm to estimate replication effect sizes and probabilities non-parametrically. We then fit parametric models (scale mixtures of two normals) that minimize the sum of squared differences with the stratified nonparametric estimates; we show that these scale mixtures of normal provide good fits the stratified nonparametric estimates of effect size and

replication probabilities for SCZ. The CM3 method does not depend on strong prior assumptions about the distribution of effect sizes, and the assumption of a scale mixture of two normal could be generalized to scale mixtures of three or more, to capture excess variation in the tails. By incorporating annotations, we show that the CM3 method results in larger numbers of identified SNPs (sensitivity) relative to the standard approach, when keeping replication rate (specificity) constant. The CM3 model may be further improved by incorporating more relevant prior information, such as gene expression, methylation, transcription regulation[10], chromatin marker annotation[28] and data about shared gene loci with other complex diseases, such as neurological disorders[8], cardiovascular disease factors[9] and immune-related diseases[29].

## Materials and Methods

### Data and Quality Control

The PGC SCZ data includes 35,476 cases and 46,839 controls[11]. Briefly, genotypes were filtered according to standard quality control parameters including: SNP missingness  $< 0.05$ , subject missingness  $< 0.02$ , and a test for deviation from Hardy-Weinberg equilibrium ( $P < 1 \times 10^{-6}$  in controls and  $1 \times 10^{-10}$  in cases). Related individuals were detected by using PLINK[30] with  $\hat{\pi} > 0.2$  with one individual from each pair removed. Principal components (PCs) were estimated using 39,239 SNPs with the program EIGENSOFT[31]. Genotype data were imputed using IMPUTE2[32] and SHAPEIT[33] based on the 1000 Genomes Project dataset. Association tests were performed on allele dosage data using the functions in PLINK with 11 PCs and study site indicators as covariates. Summary statistic p-values were generated by meta-analysis using an inverse-weighted fixed-effects model[34]. For detailed information, see the primary paper[11].

The bipolar disorder (BIP) sample consisting of the “BOMA-Bipolar”, the “Trinity College Dublin”, the “University of Edinburgh”, the “GlaxoSmithKline”, the “Systematic Treatment Enhancement Program for Bipolar Disorder”, the “University College London”, the “Thematically Organized Psychoses”, the “Wellcome Trust Case Control Consortium, WTCCC” and the “Research Program, Washington University at St. Louis, University of Pennsylvania, University of Chicago, Rush Medical School, University of Iowa, University of California, San Diego, University of California, San Francisco, and University of Michigan” studies from Sklar, *et al.*[7] were used in the current study. Genotype data were processed using the same QC parameters as for SCZ. Individuals related to or duplicated with the PGC SCZ sample were detected by PLINK with  $\hat{\pi} > 0.2$  and were removed. In total, 6,969 cases and 7,424 controls were analyzed. After QC procedures, sub-study data were combined and a mega-analysis was performed using PLINK, including the first 6 PCs and study site indicators as covariates.

### LD-informed Annotation Scores

A set of eight real-valued annotation scores, for each of the 9,266,541 SNPs analyzed here, were calculated based on the degree of correlation of the SNP with the eight different annotation categories that gave the highest genomic enrichment, as described in Schork *et al.*[5]. These categories are: exon, intron, 5' untranslated region (5'UTR), 3' untranslated region (3'UTR), 1 and 10 kilo-basepairs upstream of the gene transcription start positions, and 1 and 10 kilo-basepairs downstream of gene transcription end positions in the UCSC database. Specifically, each score was computed as the sum of LD  $r^2$  for the given SNP with, respectively, SNPs in each of the eight positional categories, with these latter SNPs comprising the full set of SNPs in the 1000 Genomes Project (approximately 39 million, the European reference sample of the November 2012 release). SNPs were assigned to non-mutually exclusive annotation categories by

thresholding the continuous category scores with an inclusive lower bound of 1.0; SNPs with scores below 1 on all functional categories were deemed *intergenic*[5]. In addition to annotation scores, the total LD (TLD) score for each SNP, given by the sum of all LD  $r^2$  for the SNP, was calculated. The correlation structure between pairs of categories is shown in [S1 Fig](#).

### Stratified Empirical Replication Effect Sizes

The 52 PGC SCZ sub-studies[11] were randomly partitioned 500 times. For each random partition, 26 of the PGC SCZ sub-studies were randomly assigned to the “discovery” sample and the complement to the “replication” sample. Inverse-variance based meta-analyses were then performed to calculate independent discovery and replication z-scores. Discovery z-scores were binned into 1,001 equally spaced intervals, and the average replication z-scores across all 500 iterations was computed for each bin. A cubic regression spline was fit to the ordinate axis (average replication z-scores) using the discovery z-score bin midpoints for the abscissa axis. This procedure was performed for all SNPs and also separately performed for strata defined by LD-weighted annotation categories ([Fig 1A](#) and [S2 Fig](#)), heterozygosity ([Fig 1B](#)), association levels with bipolar disorder ([Fig 1C](#)), and overall relative enrichment scores (described below; [Fig 3A](#) and [S3 Fig](#)).

### Relative Enrichment Score

Let  $p_i$  denote the p-value of the  $i$ th SNP from the full PGC sample. We define  $Y_{thresh,i} = 1$  if  $p_i \leq p_{thresh}$  for a pre-set threshold  $p_{thresh}$  and  $Y_{thresh,i} = 0$  otherwise. In the current study  $p_{thresh} = 10^{-3}$  was used (other choices of  $p_{thresh}$  lead to similar results, see [S1 Text](#)). A multiple logistic regression model was fit:  $\text{logit}[p_r(p_i \leq p_{thresh} | X = x_i)] = \beta x_i$ , where  $x_i$  are the values of the predictive variables for the  $i$ th SNP, i.e., annotation scores, total LD score, heterozygosity  $H = 2k(1-k)$  where  $k$  is the SNP minor allele frequency from the 1000 Genomes Project European subpopulation, and the squared z-score of the SNP association with bipolar disorder. The *relative enrichment score* for the  $i$ th SNP is defined as the estimated value of  $P_r(p_i \leq p_{thresh} | X = x_i)$  from this model. Note, before computing the relative enrichment scores, SNPs located in the extended Major Histocompatibility Complex region (xMHC, chr6: 2562429–33368333, in total 6,467 SNPs) were removed and the remainder pruned at LD  $r^2 < 0.8$ , i.e., keeping the SNP with the smallest p value in each LD block, so that in total 2,863,099 SNPs were analyzed.

### Gaussian Mixture Model

P-values from the GWAS studies were transformed into z-scores by the inverse standard normal cumulative distribution function, taking the same sign from the original study. The z-score of  $i$ th SNP can be modeled as  $z_i = \sqrt{n}\delta_i + \varepsilon_i$ , where  $n$  is the study effective sample size and  $\delta_i$  is the effect size, independent of the zero-mean Gaussian residual error term  $\varepsilon_i \sim N(0, \sigma_0^2)$ . In a commonly employed mixture model framework[13], it is assumed that some proportion  $\pi_0$  of SNPs are null ( $\delta_i = 0$ ), and the proportion  $\pi_1 = 1 - \pi_0$  are non-null ( $\delta_i \neq 0$ ) [13]. More generally, we make the assumption that an effect is “small” with prior probability  $\pi_0$  and “large” with prior probability  $\pi_1$ . The class of “small effects” includes the possibility of null effects as a special case. The small effect component is modeled by the Gaussian density  $\phi(0, nH_i\sigma_1^2)$ , and the large effect component is modeled by  $\phi(0, nH_i(\sigma_1^2 + \sigma_2^2))$ , where  $H_i = 2k_i(1-k_i)$  is the heterozygosity and  $k_i$  is the minor allele frequency of the  $i$ th SNP. The two component densities for z-scores are thus

$$f_0(z_i) = \phi(z_i | 0, \sigma_0^2 + nH_i\sigma_1^2)$$

and

$$f_1(z_i) = \phi(z_i|0, \sigma_0^2 + nH_i(\sigma_1^2 + \sigma_2^2)).$$

The unconditional marginal mixture density for z-scores is then given by

$$f(z_i) = \pi_0 f_0(z_i) + \pi_1 f_1(z_i)$$

Note, when  $\sigma_1^2 = 0$ , this reduces to the standard mixture of null (point mass at zero) and non-null (normally-distributed) z-scores.

### Posterior Effect Sizes and Predicted Replication Probabilities

Given the two component mixture model of effect sizes, the expected posterior effect size  $\delta_i$  given  $Z = z_i$  is given by [13], p. 223,

$$E\{\sqrt{n}\delta_i|Z = z_i\} = z_i + \sigma_0^2 \frac{d}{dz_i} \log\{f(z_i)\} = z_i[Afdr(z_i) + Btdr(z_i)] \tag{1}$$

$$A = \frac{nH_i\sigma_1^2}{\sigma_0^2 + nH_i\sigma_1^2}, B = \frac{nH_i(\sigma_1^2 + \sigma_2^2)}{\sigma_0^2 + nH_i(\sigma_1^2 + \sigma_2^2)},$$

where  $fdr(z_i)$  is the *local false discovery rate*, i.e., the posterior probability of a SNP being in the small effect component,

$$fdr(z_i) = \pi_0 f_0(z_i)/f(z_i). \tag{2}$$

and  $tdr(z_i) = 1 - fdr(z_i)$  is the *true discovery rate*, i.e., the posterior probability that a given SNP belongs to the large effect component.

The finite-sample *predicted replication probability* for a SNP  $i$  given the observed z-score  $Z = z_i$ , is defined as the probability that the SNP in a *de novo* replication sample, with effective sample size  $n_r$ , will have a z-score having the same sign and a magnitude equal or above certain threshold,  $z_\alpha$ . Formally,  $repl(z_i)^i = \Phi(-z_\alpha|\mu_{r,0}, \sigma_{r,0}^2)fdr(z_i) + \Phi(-z_\alpha|\mu_{r,1}, \sigma_{r,1}^2)tdr(z_i)$ , where  $\Phi$  is the Gaussian cumulative distribution function, and

$$\mu_{r,0} = -\left(\frac{\sqrt{nn_r}H_i\sigma_1^2}{\sigma_0^2 + nH_i\sigma_1^2}\right)|z_i|$$

$$\sigma_{r,0}^2 = \sigma_0^2 + n_r H_i \sigma_1^2 - \frac{nn_r(H_i\sigma_1^2)^2}{\sigma_0^2 + nH_i\sigma_1^2}$$

$$\mu_{r,1} = -\left(\frac{\sqrt{nn_r}H_i(\sigma_1^2 + \sigma_2^2)}{\sigma_0^2 + nH_i(\sigma_1^2 + \sigma_2^2)}\right)|z_i|$$

$$\sigma_{r,2}^2 = \sigma_0^2 + n_r H_i (\sigma_1^2 + \sigma_2^2) - \frac{nn_r(H_i(\sigma_1^2 + \sigma_2^2))^2}{\sigma_0^2 + nH_i(\sigma_1^2 + \sigma_2^2)}. \tag{3}$$

### Parameter Estimation

To estimate the model parameters, empirical replication effect sizes were calculated as described above, but in addition to the split-half discovery/replication breakdown of the 52 PGC sub-studies the procedure was repeated for discovery samples equal to 20%, 30%, and 40% of the total, with the complement being the replication sample. The four unknown parameters,  $\pi_1$ ,  $\sigma_0^2$ ,  $\sigma_1^2$  and  $\sigma_2^2$  from the scale-mixture of Gaussians model are then estimated by

minimizing the squared differences between the model-based and empirical (nonparametric) estimates for the posterior mean effect size and the mean of the square of the effect size.

Note, each iteration of the procedure produces an unbiased estimate of the posterior effect size means and variances, conditional on the discovery  $z$ -scores. The purpose of averaging across 500 random iterations is to smooth out the random differences present in each arbitrary partition of the sample into discovery and replication samples. Since each iteration of the sample is unbiased, the average across all iterations is again unbiased for the conditional posterior means and variances. Details of the random partitioning procedure to produce the nonparametric estimates and the quadratic estimating equations used to estimate the mixture model parameters are detailed in the [S1 Text](#).

To incorporate relative enrichment scores, SNPs are first stratified by predicted enrichment score computed from the logistic regression as described above. Nonparametric replication means and variances are computed for each stratum using the random partitioning procedure. Then, quadratic estimating equations are used to produce mixture model parameter estimates for each enrichment stratum separately. The predicted *a posteriori* effect sizes and replication probabilities are computed by Eqs (1) and (3), using the stratified parameter estimates from the quadratic estimating equations.

SNPs in the xMHC region were excluded, and the remaining 9,202,374 SNPs were randomly pruned using the LD structure from 1000 Genomes Project European subpopulation at  $r^2 < 0.8$  (see [S1 Text](#)).

### Stratified Empirical Replication Probability

Discovery and replication  $z$ -scores were computed as described above from 500 random partitions. The  $-\log_{10}$   $p$ -values computed from the discovery  $z$ -scores were binned into 1001 equally spaced bins, and the proportion of SNPs in each bin with replication  $p$ -value  $< 0.05$  was recoded as the *empirical replication rate*. The same procedure was performed on each stratum defined by predicted relative enrichment score ([Fig 3B](#) and [S4 Fig](#)). SNPs located in the xMHC region were removed and the remainder randomly pruned at LD  $r^2 < 0.8$  before performing the analysis (see [S1 Text](#) for random-pruning procedure).

### Relative Importance of Enrichment Sources

The sources of enrichment were grouped into four categories: LD-weighted genomic annotations, heterozygosity, pleiotropy with bipolar disorder (“pleiotropy” in this context is that the distribution of the summary statistics for one trait depends on those of another (“pleiotropic”) trait. No assumptions are made regarding the specific molecular, biological, or etiological factors underlying this relationship), and total LD score. Note, Then, four reduced logistic regression models were fitted. For each reduced model, one of the enrichment categories was excluded from the model, and the contribution of the deleted category was assessed by the difference in Nagelkerke’s  $R^2$  between the reduced model and the full model including all four categories. To investigate the effect of the threshold  $p_{thresh}$  used in dichotomizing the nominal  $p$ -values, this procedure was repeated with  $p_{thresh} = 10^{-2}, 10^{-4}, 10^{-5}$ . As before, SNPs located in the xMHC region were removed and then pruned at LD  $r^2 < 0.8$ , keeping the SNP with the smallest  $p$ -value in each LD block.

### Numerical Computation

See [S1 Text](#) for detailed numeric estimation of model parameters. Data QC and GWAS analysis were performed on the Genetic Cluster Computer hosted by the Dutch National Computing

and Networking Services (<http://www.geneticcluster.org/>). And the polygenic analysis was performed using PLINK[30].

## Supporting Information

### S1 Text. Supporting Methods and author lists of PGC and ENIGMA.

(PDF)

**S1 Table. Gene loci implicated by SNPs with predicted replication probability > 0.8 for Schizophrenia.** The SNPs identified by CM3 at predicted replication probabilities > 0.8 were pruned at  $LD\ r^2 < 0.1$  (after removing SNPs in the xMHC region), keeping SNP having largest predicted replication probability in each LD block, and, clumped by 250kbp. Shaded rows indicate the genomic regions having SNP above genome wide significant threshold ( $5 \times 10^{-8}$ ). The locus number (Loci), leading SNPs (LeadingSNP), reference allele(A1), chromosome numbers (Chrnum), genomic position (Pos), predicted replication probability (Pred\_Repl), p-values from the primary study (P), predicted replication probability excluding Bipolar disorder from the enrichment sources (Pred\_Repl\_noBIP) and closest genes in the region (Genes) are listed from left to right.

(XLS)

### S2 Table. Gene loci implicated by SNPs with predicted replication probability > 0.8 Crohn's disease.

The SNPs identified by CM3 at predicted replication probabilities > 0.8 were pruned at  $LD\ r^2 < 0.1$ , keeping SNP having largest predicted replication probability in each LD block, and, clumped by 250kbp. Shaded rows indicate the genomic regions having SNP above genome wide significant threshold ( $5 \times 10^{-8}$ ). The locus number (Loci), leading SNPs (LeadingSNP), reference allele(A1), chromosome numbers (Chrnum), genomic position (Pos), predicted replication probability (Pred\_Repl), p-values from the primary study (P), predicted replication probability excluding Bipolar disorder from the enrichment sources (Pred\_Repl\_noBIP) and closest genes in the region (Genes) are listed from left to right.

(XLS)

### S3 Table. Gene loci implicated by SNPs with predicted replication probability > 0.8 for Putamen volume.

The SNPs identified by CM3 at predicted replication probabilities > 0.8 were pruned at  $LD\ r^2 < 0.1$  keeping SNP having largest predicted replication probability in each LD block, and, clumped by 250kbp. Shaded rows indicate the genomic regions having SNP above genome wide significant threshold ( $5 \times 10^{-8}$ ). The locus number (Loci), leading SNPs (LeadingSNP), reference allele(A1), chromosome numbers (Chrnum), genomic position (Pos), predicted replication probability (Pred\_Repl), p-values from the primary study (P), predicted replication probability excluding Bipolar disorder from the enrichment sources (Pred\_Repl\_noBIP) and closest genes in the region (Genes) are listed from left to right.

(XLS)

**S1 Fig. Correlation between enrichment factors.** Correlations between enrichment factors. The lower triangle shows Pearson's correlation coefficient and upper triangle shows the Spearman's rank correlation. The saturation of colour encodes the magnitude. The ellipses indicate the direction and magnitude. SNPs in the extended MHC region were removed and pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ , i.e. retain SNP having the smallest p-value from the full PGC SCZ sample in each LD block.

(EPS)

**S2 Fig. Stratified mean replication z-scores by all genomic categories studied.** Mean replication z-scores for PGC SCZ SNPs from non-parametric estimates for categories stratified by

LD-weighted annotation categories. The categories include exon, intron, 5' un-translated region (5UTR), 3' un-translated region (3UTR), 10 and 1 kilo-basepair upstream of the gene transcription start positions (10kup, 1kup), 10 and 1 kilo-basepair downstream of gene transcription end positions (10kdown, 1kdown) in the UCSC database, and, SNPs with scores below 1 on all functional categories (intergenic). 26 PGC SCZ sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. All data were based on the average of 500 random draws.  
(EPS)

**S3 Fig. Mean replication z-scores stratified by enrichment scores across different re-sampling proportions.** The observed (solid lines) and predicted (dotted lines) mean z-scores in replication sample (y axis) were plotted against the z-scores in the discovery sample (x axis). The shrinkage of replication z-scores is differentiated by disjoint intervals of relative enrichment scores. **A)** 10, **B)** 17 and **C)** 20 PGC SCZ sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 10 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched stratum (Bin1) to the most enriched stratum (Bin10). All data were based on the average of 500 random draws. At each iteration the SNPs in the extended MHC region were removed and randomly pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .  
(EPS)

**S4 Fig. Mean replication probability stratified by enrichment scores across different re-sampling proportions.** The observed (solid lines) and predicted (dotted lines) replication probabilities were plotted against the negative common logarithm of nominal p-values of Schizophrenia SNPs in discovery sample (x axis). **A)** 10, **B)** 17 and **C)** 20 PGC SCZ sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 10 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched (Bin1) to the most enriched stratum (Bin10). All data were based on the average of 500 random draws. At each iteration the SNPs in the extended MHC region were removed and randomly pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .  
(EPS)

**S5 Fig. Mean replication squared z-scores stratified by enrichment scores across different re-sampling proportions.** The observed (solid lines) and predicted (dotted lines) squared mean z-scores in replication sample (y axis) were plotted against the z-scores in the discovery sample (x axis). **A)** 10, **B)** 17, **C)** 20 and **D)** 26 PGC sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 10 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched (Bin1) to the most enriched stratum (Bin10). All data were based on the average of 500 random draws. At each iteration the SNPs in the extended MHC region were removed and randomly pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .  
(EPS)

**S6 Fig. CM3 improves power of identifying gene loci by different splits.** The average empirical cumulative replication rates (y axis) are plotted against the number of SNPs replicating at that rate  $> 0.5$  (x axis), after removing MHC region SNPs and pruning at LD  $r^2 < 0.1$ . Colors indicate different sorting criteria (green: sorted by prediction replication probability and blue sorted by nominal p values). **A.** For each iteration, 11 PGC schizophrenia sub-studies were randomly assigned to the discovery sample, and the rest to the replication sample. **B.** For each iteration, 16 PGC schizophrenia sub-studies were randomly assigned to the discovery sample, and

the rest to the replication sample. **C.** For each iteration, 21 PGC schizophrenia sub-studies were randomly assigned to the discovery sample, and the rest to the replication sample. The average values over 500 iterations are shown.

(EPS)

**S7 Fig. Squared z-score as function of Heterozygosity.** Linear relationship between Heterozygosity (H) and  $z^2$  statistic. **A)** uncorrected; **B)** corrected for imputation  $R^2$ . The heterozygosity computed from the 1000 Genomes Project European population is divided into 500 equally spaced bins (x axis). Then, the mean  $z^2$  corresponding to each bin is plotted on the y-axis. The  $z^2$  and the imputation  $R^2$  for SNPs are obtained from the full PGC SCZ sample. SNPs in the extended MHC region were removed and pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ . Dotted line indicates the predicted confidence interval.

(EPS)

**S8 Fig. Ranks of SNP association with schizophrenia by CM3.** The common logarithm of the rank of SNPs based on the predicted replication probability from CM3 (y axis), using the full PGC SCZ sample, is plotted against the ranks based on p-values (x axis). The change due to sorting by CM3 is most pronounced for SNPs having smaller effect sizes. The SNPs along the red dashed line indicate no change in the ranks. The full PGC SCZ sample was used and SNPs in the extended MHC region were removed and then pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .

(EPS)

**S9 Fig. Mean replication z-score and replication rate for un-stratified data. A.)** The observed (solid lines) and predicted (dotted lines) mean z-scores in replication sample (y axis) were plotted against the z-scores in the discovery sample (x axis). **B.)** The observed (solid lines) and predicted (dotted lines) replication probabilities were plotted against the negative common logarithm of nominal p values of SNPs in discovery sample (x axis). All data were generated by randomly assigning 26 of the PGC SCZ sub-studies as discovery sample and 26 as replication sample. The averaged value over 500 iterations was shown. At each iteration the SNPs in the extended MHC region were removed and randomly pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ . The GWAS significant threshold  $p = 5 \times 10^{-8}$  ( $-\log_{10}(p) = 7.3$ ) corresponding to a predicted replication rate 0.8571.

(EPS)

**S10 Fig. Relative contribution of enrichment sources.** Explained variance by different enrichment sources. **A)** The  $-\log_{10}p$  SCZ is transformed to binary variable by different threshold (coded by colors) and the Nagelkerke's  $R^2$  is computed by subtracting from the  $R^2$  of the full model the  $R^2$  of the reduced model, namely, excluding the corresponding source. **B)** The  $z^2$  SCZ is regressed on different enrichment sources. The adjusted  $R^2$  is computed by subtracting from the  $R^2$  of the full model the  $R^2$  of the reduced model. The full PGC SCZ sample was used and SNPs in the extended MHC region were removed and pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .

(EPS)

**S11 Fig. Comparison between the CM3 and fGWAS methods by empirical replication rates.** The average empirical cumulative replication rates (y axis) are plotted against the number of SNPs replicating at that rate  $> 0.5$  (x axis), after removing MHC region SNPs and pruning at LD  $r^2 < 0.1$ . The 52 PGC schizophrenia sub-studies were randomly split into discovery and replication groups 10 times, each with 26 sub-studies. The CM3 and fGWAS methods were applied to the discovery sample, including 10k up, 1kup, 5'UTR, exon, intron, 3'UTR, 1k

down and 10k down as enrichment factors. In addition, the z-squared of the bipolar sample, heterozygosity and total LD were also included for CM3. The SNPs in the replication sample were sorted by the predicted replication probability (pred repl prob, green) and by posterior probability of association (PPA, blue) from fGWAS. The average empirical cumulative replication rates for the top 10,000 SNPs were plotted.

(EPS)

**S12 Fig. Mean replication z-scores stratified by enrichment scores across different re-sampling proportions for Crohn's disease and Putamen volume.** The observed (solid lines) and predicted (dotted lines) mean z-scores in replication sample (y axis) were plotted against the z-scores in the discovery sample (x axis). The shrinkage of replication z-scores is differentiated by disjoint intervals of relative enrichment scores. For Crohn's disease: A) 2, B) 3 and C) 4 sub-studies out of 8 sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 6 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched stratum (Bin1) to the most enriched stratum (Bin6). All data were based on the average of all possible combination of random draws. Data was genomic inflation corrected before analysis. For Putamen volume, A) 8, B) 10 and C) 13 sub-studies out of 26 sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 6 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched stratum (Bin1) to the most enriched stratum (Bin6). All data were based on the average of 100 random draws. At each iteration, the SNPs were randomly pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .

(EPS)

**S13 Fig. Mean replication probability stratified by enrichment scores across different re-sampling proportions for Crohn's disease and Putamen volume.** The observed (solid lines) and predicted (dotted lines) replication probabilities were plotted against the negative common logarithm of nominal p-values of SNPs in discovery sample (x axis). For Crohn's disease: A) 2, B) 3 and C) 4 sub-studies out of 8 sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 6 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched stratum (Bin1) to the most enriched stratum (Bin6). All data were based on the average of all possible combination of random draws. Data was genomic inflation corrected before analysis. For Putamen volume, A) 8, B) 10 and C) 13 sub-studies out of 26 sub-studies were randomly assigned as discovery sample and the remaining sub-studies as replication samples. Colors indicate the 6 disjoint intervals or bins of relative enrichment scores, ranging from the least enriched stratum (Bin1) to the most enriched stratum (Bin6). All data were based on the average of 100 random draws. At each iteration the SNPs were randomly pruned based on the 1000 Genomes Project European population at  $r^2 < 0.8$ .

(EPS)

**S14 Fig. CM3 improves power of identifying gene loci by different splits for Crohn's disease and Putamen volume.** The average empirical cumulative replication rates (y axis) are plotted against the number of SNPs replicating at that rate  $> 0.5$  (x axis), after pruning at LD  $r^2 < 0.1$ . Colors indicate different sorting criteria (green: sorted by prediction replication probability and blue sorted by nominal p values). For Crohn's disease, at each iteration A. 2, B. 3 and C. 4 sub-studies out of 8 sub-studies were randomly assigned to the discovery sample, and the rest to the replication sample. The averaged values of all possible combination were shown. For Putamen volume, at each iteration, A. 8, B. 10 C.13 sub-studies out of 26 sub-studies were

randomly assigned to the discovery sample, and the rest to the replication sample. The average values over 100 iterations are shown.

(EPS)

**S15 Fig. Relative importance of sources for enrichment measured by AUC.** The relative importance of different sources of enrichment (x axis) for explaining SNP association with schizophrenia was measured by the improvement in the areas under the receiver operating characteristic curve (improvement in AUC). The enrichment sources were: total linkage disequilibrium (TLD); the squared z-scores of SNP association with bipolar disorder (BIP); the LD weighted genomic annotation scores (Annot); and the heterozygosity (H).

(EPS)

## Acknowledgments

The authors wish to thank the participants of the studies for their contribution, as well as all researchers in Bipolar Disorder and Schizophrenia Working Groups of the Psychiatric Genomic Consortium (PGC) and the Enhancing Neuro Imaging Genetics through Meta Analysis Consortium (ENIGMA) who contributed with GWAS data. A full list of these researchers and their affiliations is provided in the [S1 Text](#).

## Author Contributions

Conceived and designed the experiments: OAA AMD. Performed the experiments: RSD AJS WKT VZ YW QC MO TW SC DRW SD CHC. Analyzed the data: YW AMD AJS WKT WL DH AW OAA AD MMN MR PMV FB. Contributed reagents/materials/analysis tools: YW AMD AJS WKT FB RSD. Wrote the paper: AD YW AMD OAA WKT AJS WL CHC RSD.

## References

1. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*. 2003; 60(12):1187–92. PMID: [14662550](#)
2. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*. 2013; 45(10):1150–9. doi: [10.1038/ng.2742](#) PMID: [23974872](#)
3. Sullivan PF. The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron*. 2010; 68(2):182–6. doi: [10.1016/j.neuron.2010.10.003](#) PMID: [20955924](#)
4. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*. 2012; 44(3):247–50. doi: [10.1038/ng.1108](#) PMID: [22344220](#)
5. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS genetics*. 2013; 9(4):e1003449. doi: [10.1371/journal.pgen.1003449](#) PMID: [23637621](#)
6. Andreassen OA, Thompson WK, Dale AM. Boosting the power of schizophrenia genetics by leveraging new statistical tools. *Schizophrenia bulletin*. 2014; 40(1):13–7. doi: [10.1093/schbul/sbt168](#) PMID: [24319118](#)
7. Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, Craddock N, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*. 2011; 43(10):977. doi: [10.1038/ng.943](#) PMID: [21926972](#)
8. Andreassen OA, Thompson WK, Schork AJ, Ripke S, Mattingsdal M, Kelsoe JR, et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics*. 2013; 9(4):e1003455. doi: [10.1371/journal.pgen.1003455](#) PMID: [23637625](#)
9. Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, O'Donovan MC, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with

- cardiovascular-disease risk factors. *The American Journal of Human Genetics*. 2013; 92(2):197–209. doi: [10.1016/j.ajhg.2013.01.001](https://doi.org/10.1016/j.ajhg.2013.01.001) PMID: [23375658](https://pubmed.ncbi.nlm.nih.gov/23375658/)
10. Kavanagh D, Dwyer S, O'Donovan M, Owen M. The ENCODE project: implications for psychiatric genetics. *Molecular psychiatry*. 2013; 18(5):540–2. doi: [10.1038/mp.2013.13](https://doi.org/10.1038/mp.2013.13) PMID: [23478746](https://pubmed.ncbi.nlm.nih.gov/23478746/)
  11. Schizophrenia Working Group of the Psychiatry Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;(511):421–7.
  12. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010; 26(17):2190–1. doi: [10.1093/bioinformatics/btq340](https://doi.org/10.1093/bioinformatics/btq340) PMID: [20616382](https://pubmed.ncbi.nlm.nih.gov/20616382/)
  13. Efron B. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*: Cambridge University Press; 2010.
  14. Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivieres S, Jahanshad N, et al. Common genetic variants influence human subcortical brain structures. *Nature*. 2015; 520(7520):224–9. doi: [10.1038/nature14101](https://doi.org/10.1038/nature14101)
  15. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*. 2010; 42(12):1118–25. doi: [10.1038/ng.717](https://doi.org/10.1038/ng.717) PMID: [21102463](https://pubmed.ncbi.nlm.nih.gov/21102463/)
  16. Zablocki RW, Schork AJ, Levine RA, Andreassen OA, Dale AM, Thompson WK. Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*. 2014; 30(15):2098–104. doi: [10.1093/bioinformatics/btu145](https://doi.org/10.1093/bioinformatics/btu145) PMID: [24711653](https://pubmed.ncbi.nlm.nih.gov/24711653/)
  17. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *The American Journal of Human Genetics*. 2012; 90(1):7–24. doi: [10.1016/j.ajhg.2011.11.029](https://doi.org/10.1016/j.ajhg.2011.11.029) PMID: [22243964](https://pubmed.ncbi.nlm.nih.gov/22243964/)
  18. Thompson WK, Wang Y, Schork AJ, Witoelar A, Zuber V, Xu S, et al. An Empirical Bayes method for estimating the distribution of effects in genome-wide association studies. *in press*.
  19. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*. 2013; 9(2):e1003264. doi: [10.1371/journal.pgen.1003264](https://doi.org/10.1371/journal.pgen.1003264) PMID: [23408905](https://pubmed.ncbi.nlm.nih.gov/23408905/)
  20. Sun L, Craiu RV, Paterson AD, Bull SB. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic epidemiology*. 2006; 30(6):519–30. PMID: [16800000](https://pubmed.ncbi.nlm.nih.gov/16800000/)
  21. Lee S-I, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, et al. Learning a prior on regulatory potential from eQTL data. *PLoS Genetics*. 2009; 5(1):e1000358. doi: [10.1371/journal.pgen.1000358](https://doi.org/10.1371/journal.pgen.1000358) PMID: [19180192](https://pubmed.ncbi.nlm.nih.gov/19180192/)
  22. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*. 2014; 94(4):559–73. doi: [10.1016/j.ajhg.2014.03.004](https://doi.org/10.1016/j.ajhg.2014.03.004) PMID: [24702953](https://pubmed.ncbi.nlm.nih.gov/24702953/)
  23. Darnell G, Duong D, Han B, Eskin E. Incorporating prior information into association studies. *Bioinformatics*. 2012; 28(12):i147–i53. doi: [10.1093/bioinformatics/bts235](https://doi.org/10.1093/bioinformatics/bts235) PMID: [22689754](https://pubmed.ncbi.nlm.nih.gov/22689754/)
  24. Eskin E. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Research*. 2008; 18(4):653–60. doi: [10.1101/gr.072785.107](https://doi.org/10.1101/gr.072785.107) PMID: [18353808](https://pubmed.ncbi.nlm.nih.gov/18353808/)
  25. Roeder K, Bacanu S-A, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics*. 2006; 78(2):243–52. PMID: [16400608](https://pubmed.ncbi.nlm.nih.gov/16400608/)
  26. Schizophrenia Psychiatry Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*. 2011; 43(10):969–76. doi: [10.1038/ng.940](https://doi.org/10.1038/ng.940) PMID: [21926974](https://pubmed.ncbi.nlm.nih.gov/21926974/)
  27. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*. 2013; 45(12):1452–8. doi: [10.1038/ng.2802](https://doi.org/10.1038/ng.2802) PMID: [24162737](https://pubmed.ncbi.nlm.nih.gov/24162737/)
  28. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*. 2013; 45(2):124–30. doi: [10.1038/ng.2504](https://doi.org/10.1038/ng.2504) PMID: [23263488](https://pubmed.ncbi.nlm.nih.gov/23263488/)
  29. Andreassen O, Harbo H, Wang Y, Thompson W, Schork A, Mattingsdal M, et al. Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Molecular psychiatry*. 2014:1–8. Epub 28 January 2014. doi: [10.1038/mp.2013.195](https://doi.org/10.1038/mp.2013.195)
  30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*. 2007; 81(3):559–75. PMID: [17701901](https://pubmed.ncbi.nlm.nih.gov/17701901/)

31. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006; 38(8):904–9. PMID: [16862161](#)
32. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*. 2011; 1(6):457–70.
33. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nature Methods*. 2012; 9(2):179–81.
34. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*. 2008; 17(R2):R122–R8. doi: [10.1093/hmg/ddn288](#) PMID: [18852200](#)