ANALYSIS

nature genetics

Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets

Zhihong Zhu[1], Futao Zhang[1], Han Hu[2], Andrew Bakshi[1], Matthew R Robinson[1], Joseph E Powell[1,3], Grant W Montgomery[4], Michael E Goddard[5,6], Naomi R Wray[1], Peter M Visscher[1,7] & Jian Yang[1,7]

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with human complex traits. However, the genes or functional DNA elements through which these variants exert their effects on the traits are often unknown. We propose a method (called SMR) that integrates summary-level data from GWAS with data from expression quantitative trait locus (eQTL) studies to identify genes whose expression levels are associated with a complex trait because of pleiotropy. We apply the method to five human complex traits using GWAS data on up to 339,224 individuals and eQTL data on 5,311 individuals, and we prioritize 126 genes (for example, *TRAF1* and *ANKRD55* for rheumatoid arthritis and *SNX19* and *NMRAL1* for schizophrenia), of which 25 genes are new candidates; 77 genes are not the nearest annotated gene to the top associated GWAS SNP. These genes provide important leads to design future functional studies to understand the mechanism whereby DNA variation leads to complex trait variation.

GWAS have identified thousands of genetic variants that are associated with diseases and traits of medical importance in humans[1]. However, the genes or DNA functional elements through which the genetic variants identified from GWAS exert their effects on diseases and traits remain largely unknown. This is mainly because of the complicated linkage disequilibrium (LD) between SNPs and causative mutations and partly because of sampling errors in test statistics. Intuitively, the genes in closest physical proximity to top associated variants are the most likely causal genes. Unfortunately, this hypothesis has not yet been tested empirically because of the lack of comprehensive follow-up functional studies for GWAS discoveries, and there have been examples in recent studies[2,3] suggesting that causal genes are distinct from the nearest genes.

A paradigmatic study[4] demonstrated the use of mRNA overexpression and suppression experiments in zebrafish to screen for the most likely causal gene at a genetic locus identified from genetic association studies in humans[5,6]. This study identified apparent differences in disease-related phenotype for individuals with the gene *kctd13* overexpressed or suppressed as compared with wild-type individuals[4]. There is an analogous natural experiment in any outbred species, including human. That is, if the expression level of a gene is influenced by a genetic variant, also known as an eQTL[7], then there will be differences in gene expression levels among individuals carrying different genotypes of the genetic variant (for example, AA, Aa and aa), analogous to the overexpression and suppression experiments (for example, AA corresponds to overexpression, Aa corresponds to wild-type expression levels and aa corresponds to suppression) (**Fig. 1**). Then, if the expression level of the gene has an effect on a trait, we will observe differences in phenotype among the different genotype groups: that is, the genetic variant will also show an effect on the trait. This approach is very similar to the concept of a Mendelian randomization (MR) analysis[8,9], where a genetic variant (for example, a SNP) is used as an instrumental variable to test for the causative effect of an exposure (for example, gene expression) on an outcome (for example, phenotype). Therefore, one can, in principle, use MR analysis to search for the most functionally relevant genes at the loci identified in GWAS for complex traits. The statistical power of an MR analysis (for example, using a two-stage least-squares approach) is proportional to the variance in outcome explained by the exposure, the variance in exposure explained by the instrument and the sample size[10,11]. Given the polygenic nature of many human complex traits[12], the variance in phenotype explained by a single genetic variant or the expression level of a single gene is likely to be very small; therefore, a very large sample size (for example, an n value on the order of tens or even hundreds of thousands of samples) is required to detect the effect of a gene on a trait using MR analysis. In practice, however, phenotype, genome-wide SNP genotype and gene expression data measured on the same samples with such a large sample size are rarely available. However, there are large amounts of summary-level data (for example, effect sizes or test statistics) from very large-scale GWAS and eQTL studies available in the public domain. In this study, we propose a method that integrates summary-level data from independent GWAS with data from eQTL studies to identify genes whose expression levels are associated with a complex trait because of pleiotropy, and we

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [2]State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou, China. [3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. [4]Molecular Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [5]Faculty of Veterinary and Agricultural Science, University of Melbourne, Parkville, Victoria, Australia. [6]Biosciences Research Division, Department of Economic Development, Jobs, Transport and Resources, Bundoora, Victoria, Australia. [7]University of Queensland Diamantina Institute, Translation Research Institute, Brisbane, Queensland, Australia. Correspondence should be addressed to J.Y. (jian.yang@uq.edu.au).

Received 5 December 2015; accepted 4 March 2016; published online 28 March 2016; doi:10.1038/ng.3538

provide an atlas of genes that are highly likely to be the functionally relevant genes underlying GWAS hits for five human complex traits and diseases.

RESULTS

Summary data–based Mendelian randomization analysis

Let y be a phenotype (outcome of interest), x be gene expression (exposure), z be a genetic variant (instrumental variable), b_{xy} be the effect size of x on y (slope of y regressed on the genetic value of x), b_{zx} be the effect of z on x, and b_{zy} be the effect of z on y. In an MR analysis, b_{xy} (defined as $b_{xy} = b_{zy}/b_{zx}$) is interpreted as the effect of x on y free of non-genetic confounders[9]. In fact, b_{xy} can also be estimated and tested using summary-level data in what we call the summary data–based Mendelian randomization (SMR) method (Online Methods). We demonstrate by simulations in the presence of a latent non-genetic confounding variable (**Supplementary Note**) that, under the assumption of either causality (where the effect of a genetic variant on a trait is mediated by gene expression) or pleiotropy (where a genetic variant has direct effects on both a trait and gene expression), SMR is equivalent to MR analysis if genotype, gene expression and phenotype data are available from the same sample and that the power of SMR can be increased by orders of magnitude if b_{zx} and b_{zy} are estimated separately from two independent samples with very large sample sizes (**Supplementary Fig. 1** and **Supplementary Table 1**). In all scenarios tested, the estimate of b_{xy} from SMR was an unbiased estimate of the effect of x on y due to a genetic factor (**Supplementary Table 2**). The simulation results also showed that there is no inflation in the SMR test statistic (T_{SMR}) if gene expression and phenotype are correlated only because of the latent non-genetic confounding variable. In addition, our simulation results showed that MR analysis (and therefore SMR) based on a single genetic variant is unable to distinguish between causality and pleiotropy regardless of whether the effect of z on x is direct or mediated by a latent variable (**Supplementary Tables 2** and **3**). Therefore, unless stated otherwise, we use 'pleiotropy' or 'pleiotropic association' to describe an association between x and y under a model of pleiotropy or causality to avoid potential misinterpretation of our method and inference about causality.

Transcriptome-wide association study using summary data

We applied SMR to test for associations between gene expression and five complex traits: height, body mass index (BMI), waist-to-hip ratio adjusted by BMI (WHRadjBMI), rheumatoid arthritis and schizophrenia (Online Methods). The estimates of SNP effects on the traits (β_{zy}) were from summary data for the latest GWAS meta-analyses for the traits[13–15] and diseases[16,17], and the estimates of SNP effects on gene expression (b_{zx}) were from the summary data of a large-scale eQTL study[18] with gene expression measured in peripheral blood (Westra eQTL data). All the data are available in the public domain (**Supplementary Table 4**). There were 14,329 probes in total in the Westra eQTL data. We included only probes with at least one *cis*-eQTL at $P_{eQTL} < 5 \times 10^{-8}$ and excluded probes in the major histocompatibility complex (MHC) region[19] because of the complexity of this region. We retained 5,967 probes for analysis. For these probes, we observed significant enrichment of the top associated *cis*-eQTLs in GWAS ($P_{enrichment} < 0.05$) for all the traits except rheumatoid arthritis (**Supplementary Table 5**). Using SMR, we tested for association between each probe and trait at the top associated *cis*-eQTL (the effect of gene expression on a trait was estimated as $\hat{b}_{xy} = \hat{b}_{zy}/\hat{\beta}_{zx}$ using the top *cis*-eQTL). This analysis can also be interpreted as a transcriptome-wide association study (TWAS) using summary data from GWAS and eQTL studies (**Fig. 2**), where b_{xy} corresponds to the association (with a positive or negative sign) between a transcript and trait free of non-genetic confounders. In total, we identified 289 genes (tagged by 327 probes) at the genome-wide significance level ($P_{SMR} < 8.4 \times 10^{-6}$) for the five complex traits and diseases (**Fig. 2** and **Table 1**), of which 202 genes passed the experiment-wise significance threshold ($P_{SMR} < 1.7 \times 10^{-6}$). In this paper, we report genes that passed the genome-wide rather than the experiment-wise significance threshold because we are interested in gene discovery for each specific trait and disease rather than the study as a whole.

Distinguishing pleiotropy from linkage

The observation of an association in an SMR test does not necessarily mean that gene expression and the trait are affected by the same underlying causal variant, as the association could possibly be due to the top associated *cis*-eQTL being in LD with two distinct causal variants, one affecting gene expression and the other affecting trait variation (**Fig. 1b**). We define this scenario as linkage, which is of less biological interest than pleiotropy. We propose a method, HEIDI (heterogeneity in dependent instruments), using multiple SNPs in a *cis*-eQTL region to distinguish pleiotropy from linkage (Online Methods). Under the hypothesis of pleiotropy, where gene expression and a trait share the same causal variant (**Fig. 1b**), the b_{xy} values calculated for any SNPs in LD with the causal variant are identical. Therefore, testing against the null hypothesis that there is a single causal variant is equivalent to testing whether there is heterogeneity in the b_{xy} values estimated for the SNPs in the *cis*-eQTL region. For each probe that passed the genome-wide significance threshold for the SMR test, we tested the heterogeneity in the b_{xy} values estimated for multiple SNPs in the *cis*-eQTL region using the HEIDI method. We detected heterogeneity for 185 of the 289 genes at $P_{HEIDI} < 0.05$. We used a *P*-value threshold of 0.05 for the HEIDI test, without correcting for multiple tests, which is conservative for gene discovery because it retains fewer genes than when correcting for multiple testing. For the remaining 104 genes (**Supplementary Table 6**) that passed the

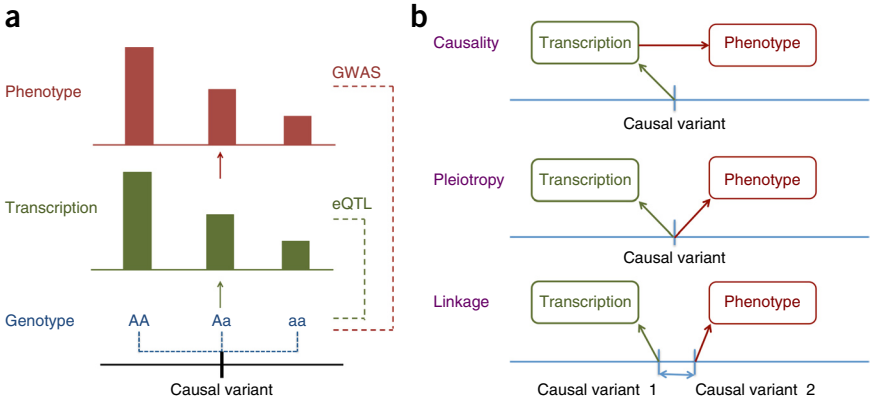

Figure 1 Association between gene expression and phenotype through genotypes. (**a**) A model of causality where a difference in phenotype is caused by a difference in genotype mediated by gene expression (transcription). (**b**) Three possible explanations for an observed association between a trait and gene expression through genotypes.

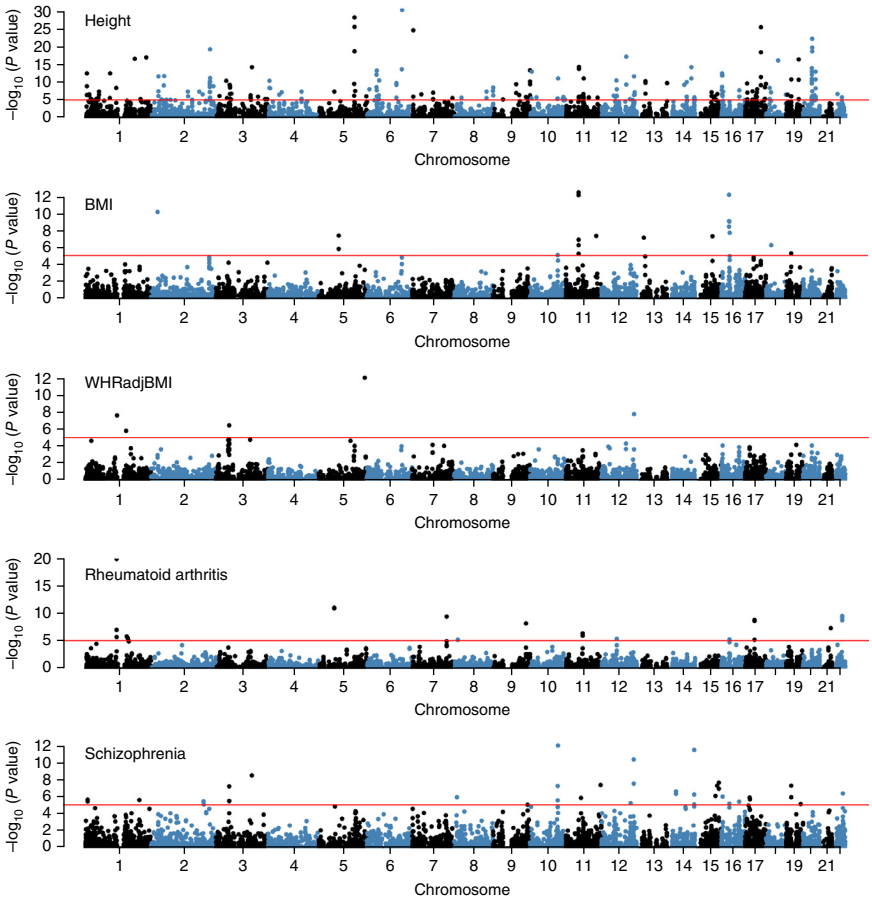

Figure 2 Manhattan plots of SMR tests for association between gene expression and complex traits. Shown on each *y* axis are the $-\log_{10}$ (*P* values) from SMR tests. The red horizontal lines represent the genome-wide significance level ($P_{SMR} = 8.4 \times 10^{-6}$).

HEIDI test ($P_{HEIDI} \geq 0.05$), we could not reject the null hypothesis that there is a single causal variant affecting both gene expression and trait variation. Hence, these 104 genes (tagged by 112 probes) are the most functionally relevant genes underlying the GWAS hits (68 for height, 9 for BMI, 2 for WHRadjBMI, 9 for rheumatoid arthritis and 16 for schizophrenia; **Table 1**) and could be prioritized in follow-up functional studies. There were three non-annotated genes. Interestingly, of the 101 annotated genes, about two-thirds (62 of 101) were not the nearest annotated gene to the top associated GWAS SNP (**Table 1**). The SMR analysis (including an SMR test followed by a HEIDI test), which only uses summary data in the public domain, provides a useful tool to prioritize genes at a known trait- or disease-associated locus for functional studies (**Fig. 3**).

Implication of new trait-associated genes

Of the 104 highly prioritized genes, 22 are new candidates: that is, there was no GWAS SNP achieving $P_{GWAS} < 5 \times 10^{-8}$ within 0.5 Mb of the probe (**Supplementary Table 7**). The GWAS signals at these gene loci probably did not reach genome-wide significance because of lack of power, despite the very large sample sizes of the GWAS (**Supplementary Table 4**). We predict that these 22 genes will be identified by GWAS or meta-analyses with even larger sample sizes in the future. This prediction is supported by evidence from the SMR analyses using data from earlier GWAS for height, BMI, WHRadjBMI, rheumatoid arthritis and schizophrenia (**Supplementary Table 4**): 30 genes passed the SMR and HEIDI tests using data from these earlier GWAS, of which 8 were 'new' (defined similarly as above), and all 8 of these genes reached genome-wide significance in the latest GWAS (**Supplementary Table 8**).

Pinpointing functionally relevant genes

TRAF1-C5 was one of the first rheumatoid arthritis–associated loci identified by GWAS[20], and the association was subsequently replicated in large-sale meta-analyses[16,21]. This locus was also found to be associated with polyarthritis in juvenile idiopathic arthritis[22]. However, it is unclear which gene is the most functionally relevant in this locus. *TRAF1* was thought more likely to be the functionally relevant gene[23] because TRAF1 protein is known to bind TNFAIP3 (*TNFAIP3* is also a rheumatoid arthritis–associated locus) and is a negative regulator of tumor necrosis factor (TNF) signaling[24,25]. This hypothesis is supported by the results from our analyses, in which the association signal from SMR (**Fig. 4**) was much stronger for *TRAF1* ($P_{SMR} = 6.2 \times 10^{-9}$) than for *C5* ($P_{SMR} = 2.4 \times 10^{-3}$) and there was heterogeneity in the b_{xy} values estimated from the SNPs in the *cis*-eQTL region for *C5* ($P_{HEIDI} = 1.2 \times 10^{-3}$) but not for those in the *TRAF1* region ($P_{HEIDI} = 0.57$), despite the eQTL association signal of the top *cis*-eQTL for *C5* ($P_{eQTL} = 2.0 \times 10^{-146}$) being much stronger than that for *TRAF1* ($P_{eQTL} = 3.8 \times 10^{-73}$). In addition, the LD r^2 value between the top GWAS SNP and the top *cis*-eQTL for *TRAF1* ($r^2 = 1.0$) is much higher than that the corresponding value for the *C5* locus ($r^2 = 0.31$). Interestingly, *TRAF1* was the only 'new' gene discovered by the SMR analysis using earlier GWAS data for rheumatoid arthritis (**Supplementary Table 8**). All these results are consistent with *TRAF1* rather than *C5* being the most functionally relevant gene in the *TRAF1-C5* locus for rheumatoid arthritis.

Tissue specificity

We used eQTL effect in blood tissue as a proxy for eQTL effect in the most relevant tissue for the trait or disease. This is not ideal because we certainly lose power for eQTLs with tissue-specific effects[26,27]. However, we gain power for genes with consistent eQTL effects across tissues because of the use of a very large sample size for eQTL analysis

Table 1 Number of genes identified by SMR for five complex human traits and diseases

Trait	Number of genes (probes) passing SMR test	Number of genes (probes) passing SMR and HEIDI tests	Nearest genes[a]
Height	212 (240)	68 (72)	28
BMI	19 (20)	9 (9)	2
WHRadjBMI	5 (5)	2 (2)	1
Rheumatoid arthritis	19 (24)	9 (12)	6
Schizophrenia	34 (38)	16 (17)	2
Total	289 (327)	104 (112)	39

[a]Number of genes corresponding to the annotated genes nearest the top associated GWAS SNPs.

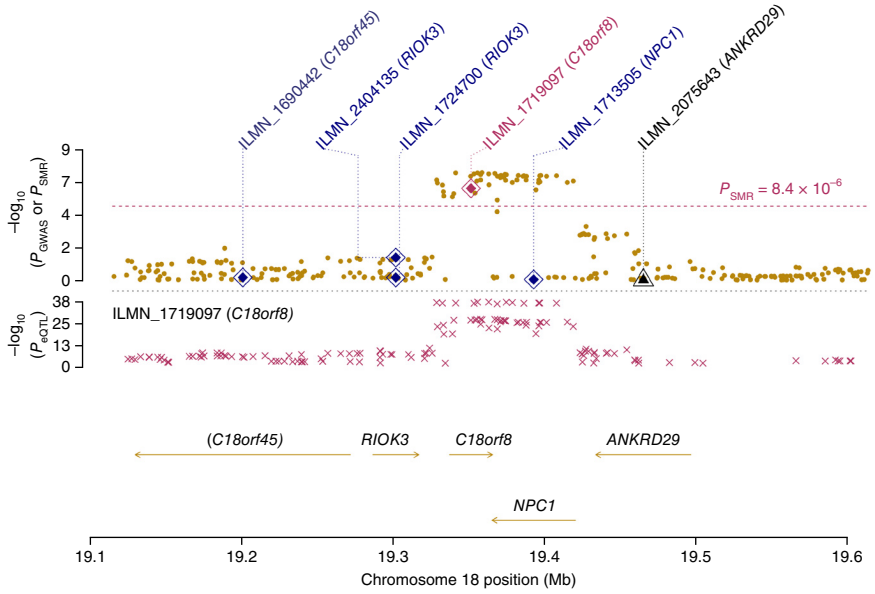

Figure 3 Prioritizing genes at a GWAS locus using SMR analysis. Shown are results at the *C18orf8* locus for BMI. Top plot, brown dots represent the *P* values for SNPs from the latest GWAS meta-analysis for BMI[14], diamonds represent the *P* values for probes from the SMR test and triangles represent probes without a *cis*-eQTL at $P_{eQTL} < 5.0 \times 10^{-8}$. Bottom plot, the eQTL *P* values of SNPs from the Westra study for the ILMN_1719097 probe tagging *C18orf8*. The top and bottom plots include all the SNPs available in the region in the GWAS and eQTL summary data, respectively, rather than only the SNPs common to both data sets. Highlighted in red is the gene (*C18orf8*) that passed the SMR and HEIDI tests.

in blood[28]. In addition, the specific tissue or cell type relevant to a trait or disease is often unknown. For example, adipose tissue was previously thought to be the most relevant tissue for BMI; however, recent studies suggest that there is an enrichment of expression for BMI-related genes in the brain[14,29]. For schizophrenia, brain tissue is seemingly the most relevant tissue, but even here the autoimmune[30] and gut-brain axis[31] hypotheses could implicate involvement of other tissues in the etiology of schizophrenia. We performed the SMR analysis for schizophrenia using eQTL data from the brain (Online Methods). There were 16 genes that passed the SMR and HEIDI tests using the Westra eQTL data in blood (**Supplementary Table 6**). We matched the transcripts of these 16 genes to brain eQTL data[32] and included in the analysis transcripts with a *cis*-eQTL at $P_{eQTL} < 1.6 \times 10^{-3}$ ($\chi^2_1 > 10$). Here we used a lower threshold for transcript inclusion than that used above because of the small sample size of the brain eQTL study and the small number of tests. Only the transcripts of six genes (one transcript for each gene) were included in the analysis (**Supplementary Table 9**). Of the six genes, two (*SNX19* and *NMRAL1*) passed both the SMR and HEIDI tests, with $P_{SMR} = 3.4 \times 10^{-5}$ and $P_{HEIDI} = 0.22$ for *SNX19* (**Fig. 5**) and $P_{SMR} = 2.2 \times 10^{-3}$ and $P_{HEIDI} = 0.39$ for *NMRAL1* (**Supplementary Fig. 2**). For both genes, the result from SMR analysis using blood eQTL data was highly consistent with that using brain eQTL data (**Supplementary Table 9**), suggesting that there is a single underlying causal variant that has pleiotropic effects on three 'phenotypes', that is, the expression level of the gene in blood, the expression level in brain and the risk of schizophrenia. We further performed SMR analysis to test for pleiotropic association between gene expression levels in blood and in the brain for each of the two genes. In this case, *x* is gene expression in blood and *y* is gene expression in brain. The results ($\hat{b}_{xy} = 1.4$, $P_{SMR} = 5.0 \times 10^{-9}$ and $P_{HEIDI} = 0.05$ for *SNX19*; $\hat{b}_{xy} = 1.4$, $P_{SMR} = 3.5 \times 10^{-5}$ and $P_{HEIDI} = 0.69$ for *NMRAL1*) are consistent with the expression level of *SNX19* and *NMRAL1* in blood being affected by the same causal variant as in the brain. As individual-level data were available for the brain eQTL study, we further used the effect sizes of the top *cis*-eQTLs for *SNX19* and *NMRAL1* estimated from the Westra data to predict the expression levels of these genes in brain. The predicted R^2 value was 21.5% for *SNX19* and 12.0% for *NMRAL1*, highly consistent with

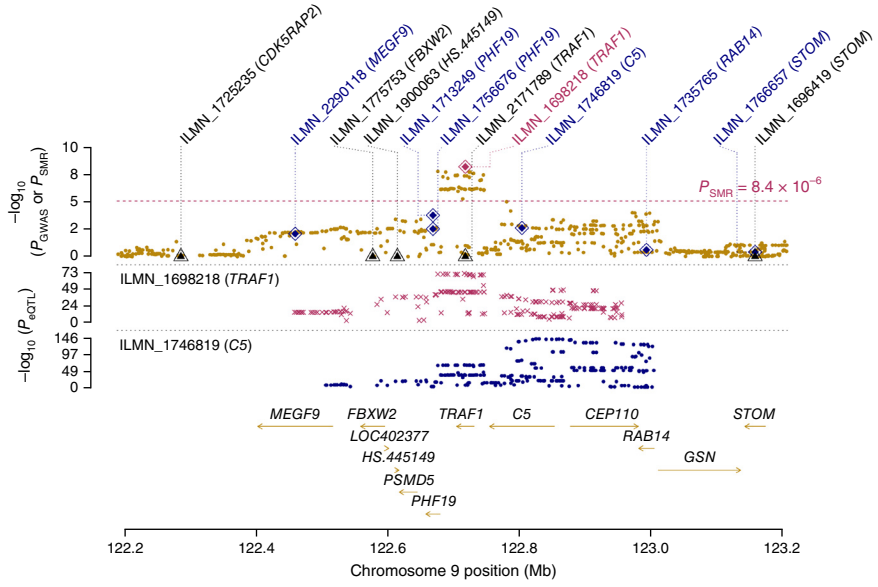

Figure 4 Prioritizing genes at the *TRAF1-C5* locus for rheumatoid arthritis. Shown are results from the SMR analysis using summary data from the latest GWAS for rheumatoid arthritis[16] and the Westra eQTL study[18]. Top plot, brown dots represent the *P* values for SNPs from GWAS, diamonds represent the *P* values for probes from the SMR test and triangles represent probes without a *cis*-eQTL at $P_{eQTL} < 5.0 \times 10^{-8}$. Bottom plot, *P* values from eQTL analysis for *TRAF1* and *C5*. Note that the difference in P_{SMR} between the two probes tagging *TRAF1* is due to the difference in *cis*-eQTL ($\hat{\beta}_{eQTL} = -0.34$, $P_{eQTL} = 3.83 \times 10^{-73}$ for ILMN_1698218 and $\hat{\beta}_{eQTL} = -0.08$, $P_{eQTL} = 8.09 \times 10^{-4}$ for ILMN_2171789) in the Westra data, consistent with low correlation of expression levels between probes ($r^2 = 0.14$) in the data (328 unrelated individuals) from the Brisbane Systems Genetics Study[39]. Plotted are all the SNPs available in the GWAS and eQTL summary data, rather than those in common to the two data sets. Probes that passed the HEIDI test ($P_{HEIDI} \geq 0.05$) are highlighted in red.

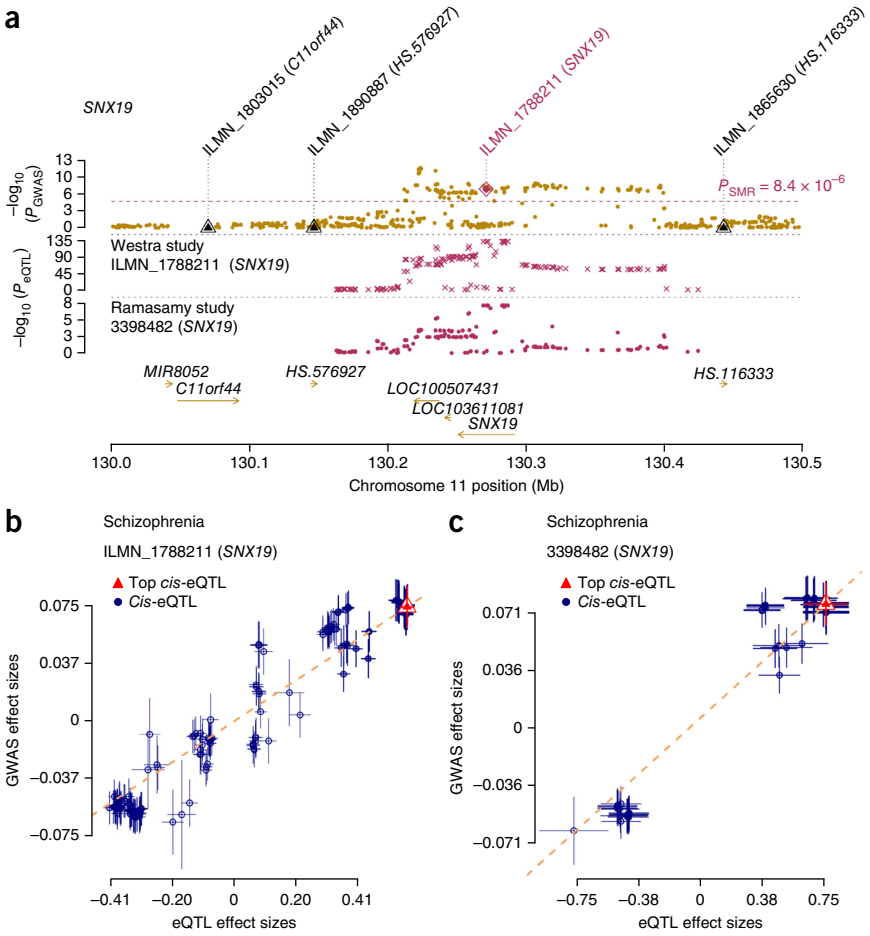

Figure 5 The *SNX19* locus for schizophrenia. (**a**) Top plot, *P* values from GWAS for schizophrenia (brown dots) and *P* value from the SMR test (diamond) using the Westra eQTL data. Bottom plot, *P* values from the Westra[18] and Ramasamy[32] eQTL studies for *SNX19*. Shown are all the SNPs available in the GWAS and eQTL data. (**b**,**c**) Effect sizes of SNPs (used for the HEIDI test) from GWAS plotted against those for SNPs from the Westra (**b**) and Ramasamy (**c**) eQTL studues. The orange dashed lines represent the estimate of b_{xy} at the top *cis*-eQTL (rather than the regression line). Error bars are the standard errors of SNP effects.

the variance in gene expression explained by the top two *cis*-eQTLs in the brain eQTL data (**Supplementary Table 9**).

Genes with multiple tagging probes

Of the 104 genes that passed the SMR and HEIDI tests, there were 19 genes tagged by multiple probes with at least one *cis*-eQTL at $P_{eQTL} < 5 \times 10^{-8}$. For six of these genes, all the tagging probes passed the SMR and HEIDI tests—*CDC16*, *CMPK1* and *PLEKHA1* for height, *ANKRD55* and *FCRL3* for rheumatoid arthritis, and *ABCB9* for schizophrenia (highlighted in bold in **Supplementary Table 6**)—providing internally consistent (although not independent) replications of the results (see **Supplementary Fig. 3** for an example). For the other 13 genes that showed different results from the SMR analysis, we tested the difference in the b_{xy} values for the tagging probes (**Supplementary Note**). We found a significant difference in b_{xy} values between the probes for nine genes after correcting for multiple testing ($P_{difference} < 0.05/13$). These differences mainly occurred because the probes tag different transcripts of the same genes (**Supplementary Table 10**). In addition, it has been suggested by a previous study that *cis*-eQTLs close to transcription end sites (TESs) tend to have exon-specific effects on gene expression[33]. We therefore tested whether the probes that showed a difference in b_{xy} (or the top associated *cis*-eQTLs of these probes) tended to be located near TESs (**Supplementary Note**). However, we observed that the mean distances of either the probes or their top associated *cis*-eQTLs to TESs were not significantly different from those obtained if the probes or *cis*-eQTLs were sampled at random (**Supplementary Table 11**).

Genetic loci with multiple association signals

We have presented above a method (HEIDI) using multiple SNPs in a *cis*-eQTL region to distinguish pleiotropy from linkage. This method assumes only one causal variant (affecting both gene expression and a trait) in the *cis*-eQTL region. Under the assumption of pleiotropy, if there are multiple causal variants in the region, the pleiotropic signal of one causal variant will be diluted by that of the other non-pleiotropic causal variants. We therefore performed GCTA conditional analysis[34], conditioning on the top associated *cis*-eQTL in both GWAS and eQTL data sets (Online Methods). If there was a secondary signal in either the GWAS or eQTL data, we performed another round of conditional analysis conditioning only on the secondary signal in both the GWAS and eQTL data sets and then reran the SMR and HEIDI tests at the top *cis*-eQTL using the estimates of SNP effects from the conditional analyses. There were 22 additional genes that passed the SMR and HEIDI tests, one of which was an unannotated gene. Of the 21 annotated genes, 3 were new (no GWAS signal within 0.5 Mb of the probe), including *KIF1B* and *ZNF318* for height and *PPP2R3C* for schizophrenia (**Supplementary Table 12**), and 15 were not the annotated genes nearest the top associated GWAS SNPs. We also attempted to perform SMR analysis for the three additional genes for schizophrenia using the brain eQTL data. However, none of these genes met our inclusion criteria for SMR analysis in the brain eQTL data because of the lack of power due to the small sample size of the brain eQTL study.

DISCUSSION

We developed our method in an MR analysis framework, with the important property that the gene-trait associations identified in the analysis are free of confounding from non-genetic factors[35]. We further proposed a method that performs a heterogeneity test (HEIDI test) to distinguish pleiotropy from linkage. There are other methods for detecting gene-trait associations using GWAS and gene expression data[36–38]. A summary description of the methods can be found in **Supplementary Table 13** and the **Supplementary Note**. There is a caveat in interpreting a *P* value from the HEIDI test. Unlike most analyses that disregard results with large statistical *P* values (not able to reject the null hypothesis that the parameter to be tested is 0), analysis using the HEIDI test disregards results with small heterogeneity test *P* values (the smaller the P_{HEIDI} value, the larger the probability of the observations being consistent with a model of linkage). It is

also important to note that a statistical analysis (for example, HEIDI or COLOC[36]) that makes use of local LD to distinguish models is unable to provide perfect separation of pleiotropy and linkage. This was demonstrated by simulations under a model of linkage where there were two distinct causal variants. In this case, the statistical power of detecting heterogeneity in b_{xy} values using the HEIDI test decreased with the LD between two causal variants (**Supplementary Fig. 4** and **Supplementary Note**). In an extreme scenario where the two causal variants are in perfect LD, the two models (pleiotropy and linkage) are undistinguishable by any statistical test. This was further demonstrated empirically with real GWAS data sets and simulated eQTL data mimicking the Westra study but with causal variants of the *cis*-eQTLs randomly placed across the genome (implicitly assuming that the effects of all the genetic variants on phenotype are functionally independent of gene expression) (**Supplementary Note**). Although for almost all the traits analyzed the observed number of probes that passed the SMR and HEIDI tests ($P_{SMR} < 5 \times 10^{-8}$ and $P_{HEIDI} \geq 0.05$) from real data analysis was significantly higher than what one would expect if *cis*-eQTLs were randomly positioned (**Supplementary Table 14**), there was a substantial number of simulated probes that passed the SMR and HEIDI tests (**Supplementary Table 14**) owing to high LD between the top GWAS SNPs and the simulated causal eQTLs (**Supplementary Figs. 4** and **5**).

In the SMR analysis of five complex traits, we initially identified associations for 289 genes by the SMR test. However, 185 of the 289 genes did not pass the subsequent HEIDI test ($P_{HEIDI} < 0.05$), suggesting that the majority of the associations identified by the SMR test could be explained by linkage due to the large number of *cis*-eQTLs widely spread across the genome. There are at least four possible reasons for the HEIDI test being too conservative (**Supplementary Note**). First, we removed probes with $P_{HEIDI} < 0.05$ without correcting for multiple testing. Second, we used summary data in the public domain that inevitably contain some errors. Third, we estimated eQTL effect sizes from *z* statistics without accounting for the difference in per-SNP sample size (**Supplementary Fig. 6**). Last but not least, multiple association signals at single loci could be detected as heterogeneity (**Supplementary Fig. 7**).

As mentioned above, to avoid any potential misinterpretation of our method and results on causality, we describe our method as an approach to identify pleiotropic associations between gene expression and a complex trait and interpret all the results under a model of pleiotropy. The SMR estimate of b_{xy}, however, has very different interpretations under different models (**Supplementary Note**). For case-control studies of diseases, we could not draw any conclusion about causality because we used eQTL effects from a population-based sample of unaffected individuals. For quantitative traits, we attempted to distinguish pleiotropy from causality with very limited data (using an additional *trans*-eQTL or multiple independent signals in a *cis* region) (**Supplementary Note**) but did not find any significant result consistent with a model of causality (**Supplementary Tables 15** and **16**). To argue that change in expression causes change in phenotype, we would need many independent SNPs that affect the expression level of the same gene from a very large eQTL study.

In conclusion, we have proposed a method to identify associations between gene expression and complex traits using summary data from GWAS and eQTL studies, followed by a heterogeneity test to distinguish pleiotropy from linkage, and we have implemented the method in a user-friendly software tool (see URLs). We observed from the analysis of five complex human traits that about two-thirds of the genes identified by SMR were not the genes nearest the top GWAS SNPs. We further demonstrated by simulations and empirical analysis that, under a model of linkage where there are two distinct causal variants, there is little statistical power to distinguish pleiotropy from linkage if the two causal variants are in high LD. Therefore, the results from SMR (or other statistical analyses) are not definitive but provide a list of prioritized genes for follow-up functional studies. In this study, we only used eQTL data from gene expression arrays, but the methods can be applied to eQTL data from RNA sequencing. Our methods can also be extended to use data from multiple-omics studies, for example, data from genome-wide associations between SNPs and DNA methylation (meQTLs), proteins (pQTLs) or metabolites (metQTLs). We can also anticipate that such multiple-omics data will also be available in multiple tissues in the near future, providing an opportunity to perform analysis in the most relevant tissues.

URLs. Blood eQTL browser, http://genenetwork.nl/bloodeqtlbrowser/; Brain eQTL Almanac (BRAINEAC), http://www.braineac.org/; annotation file for the Illumina HumanHT-12 v3.0 Gene Expression BeadChip, https://support.illumina.com/downloads/humanht-12_v3_product_files.html; SMR software, http://cnsgenomics.com/software/smr/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. McGrath for helpful comments. This research was supported by the Australian Research Council (DP130102666), the Australian National Health and Medical Research Council (grants 1078037, 1048853 and 1046880) and the Sylvia and Charles Viertel Charitable Foundation. This study makes use of data from the database of Genotypes and Phenotypes (dbGaP) available under accession phs000090.v3.p1 (see the **Supplementary Note** for the full set of acknowledgments for these data).

AUTHOR CONTRIBUTIONS

J.Y. conceived and designed the study. J.Y. and Z.Z. derived the theories. Z.Z. performed simulations and statistical analyses. F.Z., Z.Z. and J.Y. developed the software tool. H.H., A.B., M.R.R., J.E.P., G.W.M., M.E.G., N.R.W. and P.M.V. contributed by providing statistical support and/or advice on interpretation of results. J.E.P., G.W.M. and P.M.V. provided the Brisbane Systems Genetics Study data. J.Y. and Z.Z. wrote the manuscript with the participation of all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Smemo, S. *et al.* Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).
3. Claussnitzer, M. *et al. FTO* obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
4. Golzio, C. *et al. KCTD13* is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485**, 363–367 (2012).
5. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
6. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
7. Jansen, R.C. & Nap, J.P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
8. Katan, M.B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* **1**, 507–508 (1986).
9. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).

10. Freeman, G., Cowling, B.J. & Schooling, C.M. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int. J. Epidemiol.* **42**, 1157–1163 (2013).
11. Brion, M.J., Shakhbazov, K. & Visscher, P.M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**, 1497–1501 (2013).
12. Yang, J. *et al.* Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLoS Genet.* **9**, e1003355 (2013).
13. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
14. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
15. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
16. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
17. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
18. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
19. Patsopoulos, N.A. *et al.* Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects. *PLoS Genet.* **9**, e1003926 (2013).
20. Plenge, R.M. *et al. TRAF1-C5* as a risk locus for rheumatoid arthritis—a genomewide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
21. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
22. Albers, H.M. *et al.* The *TRAF1/C5* region is a risk factor for polyarthritis in juvenile idiopathic arthritis. *Ann. Rheum. Dis.* **67**, 1578–1580 (2008).
23. Xavier, R.J. & Rioux, J.D. Genome-wide association studies: a new window into immune-mediated diseases. *Nat. Rev. Immunol.* **8**, 631–643 (2008).
24. Tsitsikov, E.N. *et al.* TRAF1 is a negative regulator of TNF signaling. enhanced TNF signaling in TRAF1-deficient mice. *Immunity* **15**, 647–657 (2001).
25. Chung, J.Y., Park, Y.C., Ye, H. & Wu, H. All TRAFs are not created equal: common and distinct molecular mechanisms of TRAF-mediated signal transduction. *J. Cell Sci.* **115**, 679–688 (2002).
26. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
27. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
28. McKenzie, M., Henders, A.K., Caracella, A., Wray, N.R. & Powell, J.E. Overlap of expression quantitative trait loci (eQTL) in human brain and blood. *BMC Med. Genomics* **7**, 31 (2014).
29. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
30. Eaton, W.W. *et al.* Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *Am. J. Psychiatry* **163**, 521–528 (2006).
31. Nemani, K., Hosseini Ghomi, R., McCormick, B. & Fan, X. Schizophrenia and the gut-brain axis. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **56**, 155–160 (2015).
32. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418–1428 (2014).
33. Veyrieras, J.B. *et al.* Exon-specific QTLs skew the inferred distribution of expression QTLs detected using gene expression array data. *PLoS One* **7**, e30629 (2012).
34. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1–S3 (2012).
35. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
36. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
37. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
38. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
39. Powell, J.E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One* **7**, e35430 (2012).

ONLINE METHODS

Summary data–based Mendelian randomization analysis. Theories and methods for MR analysis have been well established[9,40]. MR analysis uses a genetic variant as an instrumental variable to estimate and test for the causative effect of an exposure variable (for example, high-density lipoprotein cholesterol) on an outcome (for example, coronary heart disease)[41]. We adopt the MR approach to test for pleiotropic association between the expression level of a gene (exposure) and a trait (outcome). We define a 'pleiotropic association' as association between a trait and gene expression due to either pleiotropy (both gene expression and the trait are affected by the same causal variant) or causality (the effect of a causal variant on the trait is mediated by gene expression). This is because the MR approach using a single genetic variant is unable to distinguish between pleiotropy and causality[35]. If we denote z as a genetic variant (for example, SNP), x as the expression level of a gene and y as the trait, then the two-step least-squares (2SLS) estimate of the effect of x on y from an MR analysis is

$$\hat{b}_{xy} = \hat{b}_{zy} / \hat{b}_{zx} \quad (1)$$

where $\hat{b}_{zy}$ and $\hat{b}_{zx}$ are the least-squares estimates of y and x on z, respectively, and b_{xy} is interpreted as the effect size of x on y free of confounding from non-genetic factors[9]. The sampling variance of the 2SLS estimate of b_{xy} is

$$\mathrm{var}(\hat{b}_{xy}) = \left[\mathrm{var}(y)(1-R^2_{xy})\right] \Big/ \left[n\,\mathrm{var}(x)R^2_{zx}\right] \quad (2)$$

where n is the sample size, R^2_{zy} is the proportion of variance in y explained by x and R^2_{zx} is the proportion of variance in x explained by z. We therefore can have a statistic

$$T_{\mathrm{MR}} = \hat{b}^2_{xy} / \mathrm{var}(\hat{b}_{xy})$$

to test the significance of b_{xy}, where $T_{\mathrm{MR}} = \chi^2_1$. This analysis, however, requires genotype, gene expression and phenotype to be measured on the same sample and the availability of individual-level data, which limits power because the non-centrality parameter (NCP) of T_{MR} is proportional to n (ref. 10) and it is currently (and in the foreseeable near future) almost unrealistic to collect genome-wide SNP genotype and gene expression data in a very large sample (for example, where there are hundreds of thousands of samples).

It has been suggested that the power of detecting b_{xy} can be greatly increased using a two-sample MR analysis[42,43]. There have been large-scale meta-analyses of GWAS for a number of complex traits and diseases with summary data (for example, estimate of effect size and its standard error for each SNP) available in the public domain. Although gene expression data are usually not available in GWAS samples, the effect of a SNP on gene expression can be estimated from an independent eQTL study of very large sample size. If the GWAS and eQTL study samples are from the same population, $\hat{\beta}_{zx}$ should also be an unbiased estimate of b_{zx}. We therefore have

$$\hat{b}_{xy} = \hat{b}_{zy} / \hat{\beta}_{zx} \quad (3)$$

where $\hat{b}_{zy}$ is the estimate of a SNP effect from a GWAS for a trait and $\hat{\beta}_{zx}$ is the estimate of a SNP effect on the expression level of a gene from an independent eQTL study. The sampling variance of $\hat{b}_{xy}$ can be calculated approximately by the Delta method[44]

$$\mathrm{var}(\hat{b}_{xy}) \approx \frac{b^2_{zy}}{\beta^2_{zx}}\left[\frac{\mathrm{var}(\hat{\beta}_{zx})}{\beta^2_{zx}} + \frac{\mathrm{var}(\hat{b}_{zy})}{b^2_{zy}} - \frac{2\,\mathrm{cov}(\hat{\beta}_{zx},\hat{b}_{zy})}{\beta_{zx}b_{zy}}\right] \quad (4)$$

where $\mathrm{cov}(\hat{\beta}_{zx},\hat{b}_{zy})$ is 0 if β_{zx} and b_{zy} are estimated from independent samples (it is also 0 if x and y are independent even though β_{zx} and b_{zy} are estimated from the same sample). The expected values of β_{zx} and b_{zy} are unknown. In practice, we can replace them by their estimates, yielding an approximate χ^2 test statistic of

$$T_{\mathrm{SMR}} = \hat{b}^2_{xy} / \mathrm{var}(\hat{b}_{xy}) \approx \frac{z^2_{zy}z^2_{zx}}{z^2_{zy} + z^2_{zx}} \quad (5)$$

where z_{zy} and z_{zx} are the z statistics from the GWAS and eQTL study, respectively.

Transcriptome-wide association study for complex traits using GWAS and eQTL summary data. We applied the SMR method to test for the association between a trait and the expression level of each gene across the whole genome using summary data from GWAS (**Supplementary Table 4**) and eQTL studies (see URLs). The eQTL summary data were from the study by Westra *et al.*[18], an eQTL meta-analysis of 5,311 samples from peripheral blood, with gene expression data observed from Illumina gene expression arrays and SNP genotype data imputed to the HapMap 2 reference panels[45]. Information about the physical positions of the probes and genes that the probes tag is from the annotation file provided by Illumina (see URLs), with the physical positions of the genes derived from Ensembl (hg18). The summary data consist of z statistics (z_{zx}) of 923,021 *cis*-eQTLs (<250 kb away from the probe) for 14,329 gene expression probes and 4,732 *trans*-eQTLs (>5 Mb away from the probe) for 2,612 gene expression probes, selected at an FDR of 0.05. Because the effect sizes of eQTLs were not available in the summary data, we estimated b_{zx} from the z statistic using the following equation (**Supplementary Note**)

$$\hat{b}_{zx} = z_{zx}S_{zx} \quad (6)$$

where $S_{zx} = 1/\sqrt{2p(1-p)(n+z^2_{zx})}$, p is allele frequency and n is sample size. Allele frequency was also not available in the summary data and was estimated from the reference sample, that is, the HapMap 2–imputed Atherosclerosis Risk in Communities (ARIC) data[46] from dbGaP. We included in the analysis only probes for which the P value of the top associated *cis*-eQTL was $<5 \times 10^{-8}$. This is because one of the basic assumptions for an MR analysis is that the instrumental variable has a strong effect on exposure, that is, the SNP is strongly associated with gene expression. We did not include the MHC region[19] (26.2–33.8 Mb on chromosome 6 based on hg18) in the analysis because of the complexity of this region. We further removed SNPs (eQTLs) with minor allele frequency (MAF) <0.01 (MAF estimated from the ARIC data). After filtering, there were 5,967 probes and 757,479 SNPs.

The GWAS summary data were from the latest meta-analyses of height[13], BMI[14], WHRadjBMI[15], rheumatoid arthritis[16] and schizophrenia[17]. The sample sizes of these studies are listed in **Supplementary Table 4**. The number of SNPs varied from 2.5 to 9.4 million across quantitative traits or diseases. We included in the analysis SNPs with MAF ≥0.01 (based on the reported MAF in the summary data) and those in common with the Westra eQTL data. For the traits (rheumatoid arthritis and schizophrenia) without reported allele frequency, we estimated allele frequency from the ARIC data. After filtering, we retained ~500,000 SNPs for analysis.

For each gene expression probe, we used the SMR method to test for the association between a trait and probe (b_{xy}) using the effect size of the top associated *cis*-eQTL from the eQTL study (β_{zx}) and the effect size of the same SNP from the GWAS (b_{zy}). We have shown by simulations (**Supplementary Fig. 1**) that the SMR test statistic is not inflated under the null (where x and y are associated only because of a latent non-genetic confounding variable). To control the genome-wide type I error rate, we used Bonferroni correction to account for multiple testing, which resulted in a genome-wide significance level of $P = 8.4 \times 10^{-6}$ (= 0.05/5,967) for each trait or disease and an experiment-wise significance level of $P = 1.7 \times 10^{-6}$ (= 8.4×10^{-6}/5).

Distinguishing functional association from linkage. We describe below a method to test for heterogeneity in dependent instruments (HEIDI). If a trait and gene expression are affected by the same causal variant (pleiotropy), then b_{xy} calculated using any SNP in LD with the causal variant is identical. This is because, under Hardy-Weinberg equilibrium, for any SNP i

$$b_{xy(i)} = \frac{b_{zy(i)}}{\beta_{zx(i)}} = \frac{b_{zy(0)}r_{0i}\sqrt{h_0/h_i}}{\beta_{zx(0)}r_{0i}\sqrt{h_0/h_i}} = \frac{b_{zy(0)}}{\beta_{zx(0)}} = b_{xy(0)} \quad (7)$$

where a subscript "0" represents the causal variant, a subscript "i" represents a SNP i, r_{0i} is the LD correlation between the causal variant and SNP i, and $h = 2p(1-p)$ with p being allele frequency. Therefore, testing linkage (two distinct causal variants, one affecting gene expression and one affecting a trait) against pleiotropy is equivalent to testing whether there is a difference between b_{xy} estimated using the top associated *cis*-eQTL ($b_{xy(\mathrm{top})}$) and that using any

doi:10.1038/ng.3538

other significant SNP in the *cis*-eQTL region ($b_{xy(i)}$). If we define $d_i = b_{xy(i)} - b_{xy(\text{top})}$, it is further equivalent to testing whether $d_i = 0$ for all the significant *cis*-eQTLs (excluding the top *cis*-eQTL). If we define $\hat{\mathbf{d}} = \{\hat{d}_1, \hat{d}_2, \ldots \hat{d}_m\}$ with m being the number of significant eQTLs (excluding the top *cis*-eQTL), we have $\hat{\mathbf{d}} \sim \text{MVN}(\mathbf{d}, \mathbf{V})$, where $\mathbf{V}$ is the (co)variance matrix with the ijth element being

$$\begin{aligned}\text{cov}(\hat{d}_i, \hat{d}_j) = \text{cov}(\hat{b}_{xy(i)}, \hat{b}_{xy(j)}) - \text{cov}(\hat{b}_{xy(i)}, \hat{b}_{xy(\text{top})}) \\ - \text{cov}(\hat{b}_{xy(j)}, \hat{b}_{xy(\text{top})}) + \text{var}(\hat{b}_{xy(\text{top})})\end{aligned} \quad (8)$$

The covariance between $\hat{b}_{xy(i)}$ and $\hat{b}_{xy(\text{j})}$ can be calculated as (**Supplementary Note**)

$$\begin{aligned}\text{cov}(\hat{b}_{xy(i)}, \hat{b}_{xy(j)}) = \frac{r_{ij}}{\beta_{zx(i)}\beta_{zx(j)}} \sqrt{\text{var}(\hat{b}_{zy(i)})\text{var}(\hat{b}_{zy(j)})} \\ + b_{xy(i)}b_{xy(j)}\left(\frac{r_{ij}}{z_{zx(i)}z_{zx(j)}} - \frac{1}{z^2_{zx(i)}z^2_{zx(j)}}\right)\end{aligned} \quad (9)$$

where r_{ij} is the LD correlation between SNPs i and j. Under the null hypothesis that there is no heterogeneity, that is, where $\mathbf{d} = 0$, we have a vector of standard normal variables $\mathbf{z}_d = \{z_{d(1)}, z_{d(2)}, \ldots, z_{d(m)}\}$ with $z_{d(i)} = \hat{d}_i / \sqrt{\text{var}(\hat{d}_i)}$ and $\mathbf{z}_d \sim \text{MVN}(0, \mathbf{R})$, where $\mathbf{R}$ is the correlation matrix with the ijth element being $r(z_{d(i)}, z_{d(j)}) = \text{cov}(\hat{d}_i, \hat{d}_j) / \sqrt{\text{var}(\hat{d}_i)\text{var}(\hat{d}_j)}$. To test against $\mathbf{d} = 0$, we construct a HEIDI test statistic $T_{\text{HEIDI}} = \mathbf{z}_d \mathbf{I} \mathbf{z}_d^T$, that is, $T_{\text{HEIDI}} = \sum_i^m z^2_{d(i)}$, with $\mathbf{I}$ being an identity matrix. This is a quadric form of standard normal variables, the distribution of which is not explicit but can be approximated by the Satterthwaite or Saddlepoint method[47,48]. In the analyses, we excluded SNPs in LD with the top *cis*-eQTL at $r^2 > 0.9$ because SNPs in almost perfect LD with the top *cis*-eQTL are not informative for the HEIDI test. We also removed SNPs in the *cis*-eQTL regions with $P_{\text{eQTL}} > 1.6 \times 10^{-3}$ (equivalent to $\chi^2_1 < 10$) to avoid weak instrumental variables[35]. We used the HapMap 2–imputed ARIC data as the reference sample to estimate the LD correlation between SNPs.

SMR analysis for schizophrenia with eQTL data from brain samples. For the 16 highly prioritized genes for schizophrenia from SMR analysis using the Westra eQTL data (blood tissue), we performed SMR analysis using the Ramasamy eQTL data (brain tissue)[32]. The individual-level genotype and gene expression data are available in the public domain (see URLs) and comprise 26,493 gene expression probes in ten brain regions and 5,878,211 genetic variants (genotyped SNPs imputed to the 1000 Genome Project reference panels[49]) on up to 134 individuals. To maximize statistical power, we used the eQTL effects estimated from gene expression averaged across all ten brain regions. Because the two studies were based on different gene expression platforms (the Westra study on Illumina and the Ramasamy study on Affymetrix), we matched the data by transcripts rather than probes. We included in the analysis only probes that had a *cis*-eQTL at $P < 1.6 \times 10^{-3}$ (equivalent to $\chi^2_1 > 10$) in the brain eQTL data and only the SNPs in common between the two data sets.

Conditional association analysis. We performed GCTA conditional analysis[34,50] using the Westra eQTL data and LD between SNPs from the HapMap 2–imputed ARIC data. The conditional analysis (conditioning on the top *cis*-eQTL) was only performed in *cis*-eQTL regions (±250 kb) of the 215 probes that passed the SMR test but failed to pass the HEIDI test. We also performed the conditional analysis using GWAS summary data of the same set of SNPs (SNPs in the *cis*-eQTL regions conditioning on the top *cis*-eQTL) for each of the five complex traits. For any of these regions where there was evidence of a secondary signal ($P_{\text{conditional}} < 5 \times 10^{-8}$) in either eQTL or GWAS data, we reran the conditional analyses in both eQTL and GWAS data conditioning on the secondary signal and then used the conditional results for both SMR and HEIDI tests.

40. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23** R1, R89–R98 (2014).
41. Ference, B.A. *et al.* Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *J. Am. Coll. Cardiol.* **60**, 2631–2639 (2012).
42. Pierce, B.L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).
43. Inoue, A. & Solon, G. Two-sample instrumental variables estimators. *Rev. Econ. Stat.* **92**, 557–561 (2010).
44. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, 1998).
45. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
46. Psaty, B.M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ. Cardiovasc. Genet.* **2**, 73–80 (2009).
47. Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935 (1999).
48. Davies, R.B. Numerical inversion of a characteristic function. *Biometrika* **60**, 415–417 (1973).
49. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
50. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).