Published in final edited form as:
*Nat Rev Genet*. 2013 July ; 14(7): 483–495. doi:10.1038/nrg3461.

# Pleiotropy in complex traits: challenges and strategies

**Nadia Solovieff**[1,2,3], **Chris Cotsapas**[4,5], **Phil H. Lee**[1,2,3], **Shaun M. Purcell**[1,2,3,6], and **Jordan W. Smoller**[1,2,3]

[1]Center for Human Genetics Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA

[2]Department of Psychiatry, Harvard Medical School, 2 West, Room 305, 401 Park Drive, Boston, Massachusetts 02215, USA

[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA

[4]Departments of Neurology and Genetics, Yale University School of Medicine, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA

[5]Medical and Population Genetics, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA

[6]Division of Psychiatric Genomics, Mount Sinai School of Medicine, 1 Gustave L. Levy Place, New York, New York 10029–6574, USA

## Abstract

Genome-wide association studies have identified many variants that each affects multiple traits, particularly across autoimmune diseases, cancers and neuropsychiatric disorders, suggesting that pleiotropic effects on human complex traits may be widespread. However, systematic detection of such effects is challenging and requires new methodologies and frameworks for interpreting cross-phenotype results. In this Review, we discuss the evidence for pleiotropy in contemporary genetic mapping studies, new and established analytical approaches to identifying pleiotropic effects, sources of spurious cross-phenotype effects and study design considerations. We also outline the molecular and clinical implications of such findings and discuss future directions of research.



Correspondence to J.W.S. jsmoller@hms.harvard.edu.

**Competing interests statement**
The authors declare no competing financial interests.

**FURTHER INFORMATION**
**Chris Cotsapas's homepage:** http://www.cotsapaslab.info
**ENCODE Project**: http://www.nature.com/encode
**Genotype-Tissue Expression eQTL Browser**: http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi
**NHGRI GWAS Catalogue**: http://www.genome.gov/gwastudies
**Online Mendelian Inheriatnce in Man (OMIM)**: http://omim.org
**PAGE network**: http://www.pagestudy.org
**Pathguide**: http://www.pathguide.org
**POLYPHEN**: http://genetics.bwh.harvard.edu/pph2/index.shtml
**SIFT**: http://sift.jcvi.org
**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**

In the past 7 years, a wave of genome-wide association studies (GWASs) has identified more than 8,500 genome-wide-significant associations with more than 350 human complex traits, including susceptibility to a wide variety of diseases[1]. An interesting observation has been that many genetic loci appear to harbour variants that are associated with multiple, sometimes seemingly distinct traits, and such associations are termed cross-phenotype (CP) associations. CP associations have been identified in several disease areas. Examples include: protein tyrosine phosphatase non-receptor type 22 (*PTPN22*) for immune-related disorders, such as rheumatoid arthritis[2], Crohn's disease[3], systemic lupus erythematosus[4] and type 1 diabetes[5]; the telomerase reverse transcriptase (*TERT*)–CLPTM1-like (*CLPTM1L*) locus for glioma, bladder and lung cancers[6]; and calcium channel, voltage-dependent, L-type, alpha 1C subunit (*CACNA1C*) for bipolar disorder and schizophrenia[7]. These CP associations highlight that these traits share common genetic pathways and underscore the relevance of pleiotropy[8,9] in human complex disease. The distinction between a CP association and pleiotropy is important to define. A CP association occurs when a genetic locus is associated with more than one trait in a study, regardless of the underlying cause for the observed association. Pleiotropy occurs when a genetic locus truly affects more than one trait and is one possible underlying cause for an observed CP association (others are discussed below).

CP effects in GWASs mirror epidemiological observations of shared heritability and comorbidity. For example, twin and family studies have long provided evidence for genetic correlations among diseases (such as major depressive disorder and generalized anxiety disorder[10], or rheumatoid arthritis and systemic lupus erythematosus[11]), suggesting a role for pleiotropic genetic effects. In addition, the co-occurrence of multiple diseases in the same individual (for example, type 1 diabetes and autoimmune thyroid disease[12]) also point to shared genetic causes.

In some cases, the same variants show association with multiple traits; in other cases, although the same overall region is implicated, distinct nearby markers show signals of association with different traits. Distinguishing the associations that represent genuinely shared effects of single variants from those that represent the effects of colocalizing but independent variants is crucial, as they imply different notions of pleiotropy and mechanistic models of shared function. In this article, we define three types of such CP genetic effects that occur when a genetic variant or gene is correlated with more than one trait: biological pleiotropy, mediated pleiotropy and spurious pleiotropy. In brief, biological pleiotropy refers to a genetic variant or gene that has a direct biological influence on more than one phenotypic trait. Mediated pleiotropy occurs when one phenotype is itself causally related to a second phenotype so that a variant associated with the first phenotype is indirectly associated with the second. Spurious pleiotropy encompasses various sources of bias that cause a genetic variant falsely to appear to be associated with multiple phenotypes.

Here, we first review evidence of CP associations in the literature and the underlying causal models that they imply. We next outline the analytical strategies that are required for detecting CP effects, particularly methods that can be readily applied to existing GWAS data sets, and how the types of pleiotropy can be distinguished and functionally characterized. Finally, we discuss the clinical implications of CP associations and visions for the future.

Overall, we conclude that despite various conceptual and technical challenges, the identification and characterization of this widespread pleiotropy is crucial for a comprehensive biological understanding of complex traits and disease states.

## Cross-phenotype effects in GWASs

The results of GWASs have highlighted numerous CP effects, particularly across autoimmune diseases and psychiatric traits (TABLE 1). Such observations have usually been incidental, and studies of different traits have independently led to discoveries of associations with the same marker or region. As the power of most GWASs is sufficient to detect only a subset of the many true associations, the chance of two independent studies both detecting a true association at the same locus is correspondingly low. Estimates of overlaps are thus likely to be conservative. Nonetheless, a startling level of overlap has been observed.

A recent evaluation of genome-wide-significant `single-nucleotide polymorphisms` (SNPs) listed in the National Human Genome Research Institute (NHGRI) Catalogue of Published Genome-Wide Association Studies found that 4.6% of SNPs and 16.9% of genes have CP effects[13]. These are underestimates as they rely on highly conservative criteria (for example, an association of genome-wide significance for each trait) and were limited by the incomplete database of GWAS-associated SNPs at the start of 2011. The first examples of cross-disease metaanalyses (using methods described later) have discovered even higher levels of overlap: Cotsapas *et al.*[14] estimate that at least 44% of SNPs associated with one autoimmune disease are associated with another. Interestingly, Sirota *et al.*[15] show that opposite effects — in which an allele appears to increase the risk of one disease trait and decrease the risk of another disease trait — are also frequent. Recently, a large meta-analysis of Crohn's disease and ulcerative colitis identified 110 SNPs that are associated with both disorders and found that 70% of SNPs were shared across other immune-mediated diseases[16].

A CP association can be observed for an individual SNP or at the level of a gene or region (including in noncoding DNA), in which different independent variants in the same gene or region affect multiple phenotypes. Both SNP-level and gene- or region-level CP effects can be considered to be real forms of pleiotropy and provide insight into the shared underlying biology. For example, variants in intron 1 of fat mass and obesity associated (*FTO*) have been robustly associated with body mass index (BMI)[17]. Recently, variants elsewhere in the gene (and not in apparent `linkage disequilibrium` with the obesity-associated SNPs) have been associated with melanoma and not with BMI[18]. CP effects outside protein-coding genes include the 9q21.3 locus[19–22] and rs6983267 (REFS 23,24; TABLE 1) and point to possible *cis*regulatory effects on gene expression[25]. In fact, the 88% of SNPs reported in the NGHRI catalogue are intronic or intergenic[1]. GWASs and other genomic analyses have also identified rare structural variations that have CP effects. For example, rare `copy number variants` (CNVs) in multiple chromosomal regions have been found to increase the risk of a range of neurodevelopmental disorders[26,27]. Distinguishing between biological and spurious pleiotropy for CNVs is particularly challenging because it is unclear whether the

same gene affects multiple traits (for biological pleiotropy) or whether different genes within the region affect different traits (for spurious pleiotropy).

Finally, studies using aggregate measures of genetic variation (such as `polygenic` genetic risk scores) have been used to demonstrate genetic covariation between two or more disorders. For example, using molecular genetic data, Purcell *et al*.[28] showed that a substantial proportion of heritability is shared between schizophrenia and bipolar disorder, which is consistent with family-based epidemiological studies[29].

## Biological pleiotropy

Characterizing the underlying biological mechanism of a pleiotropic effect is a major challenge in the field as many alternative models for an apparent CP effect can fit the observed data (FIG. 1). Pleiotropy can occur at the allelic level, where a single causal variant is related to multiple phenotypes (FIG. 1a), or at the gene (or region) level, at which multiple variants in the same gene (or region) are associated with different phenotypes (FIG. 1b,c). For example, the common coding variant in *PTPN22* described above seems to influence the function of various subpopulations of T cells[30] but also interferes with the removal of auto-reactive B cells[31]. The equivalent variant in mice promotes degradation of LYP (also known as PEP), which is the protein encoded by *PTPN22*. This suggests that this is a loss-of-function allele[32], although much more work is required to demonstrate the causal mechanism[33]. This variant decreases the risk of Crohn's disease but increases the risk of rheumatoid arthritis and type 1 diabetes[34], prompting questions about whether the opposite effects correspond to functional changes in different cells or whether the overall homeostatic changes to T and/or B cell populations are responsible for risk versus protective states. At first glance, several scenarios fit these observations: distinct effects of the same allele in different cell populations underlying associations with different diseases or disease groups; a single molecular effect having multiple morphological or physiological consequences; or a CP effect tagging two different causal variants within the same gene (FIG. 1b) that result in different functions and affect different phenotypes.

An example of biological pleiotropy in an intergenic region is the rs6983267 SNP on chromosome 8q24 that is a risk variant for prostate and colorectal cancer (TABLE 1). This allele alters the ability of this region to act as an enhancer for the downstream *MYC* oncogene in both colon and prostate tissue types[35,36].

## Mediated pleiotropy

CP effects can also occur when one phenotype is causal for a second phenotype and a genetic variant is directly (or 'more proximally') associated with the first phenotype (FIG. 1d). In such cases of mediated pleiotropy, the genetic variant will be associated with both phenotypes if tested separately. Mediated pleiotropy is a real form of pleiotropy, in contrast to spurious pleiotropy, but it is important to distinguish this category from what we call biological pleiotropy in order to describe the underlying aetiology of the phenotypes properly. For example, genetic variants have been found to be associated with both low-density lipoprotein (LDL) levels and risk of myocardial infarction[37]. However, LDL levels are themselves risk factors for myocardial infarction, so we must deconvolute whether a

genetic variant influences myocardial infarction risk by altering LDL levels or whether it has an additional effect that is independent of LDL levels. Another example includes the observed association of 15q24–15q25.1 with lung cancer[38] and nicotine dependence[39], which has spurred a debate about whether this region has a direct effect on lung cancer[40].

## Sources of spurious pleiotropy

There are several ways in which a spurious CP association can occur and falsely suggest underlying pleiotropy. These include defects in the studies that identify CP effects, such as `ascertainment bias`, phenotypic misclassification and shared controls (FIG. 1e). Further details on these aspects and their minimization by careful study design are described in BOX 1.

Additionally, spurious associations can arise when there is ambiguity in mapping the true underlying causal variant. There is currently limited evidence that the primary SNPs identified in GWASs are causal variants; instead, they are often `tag SNPs` that typically associate with the trait because they are in strong linkage disequilibrium (LD) with the nearby causal variant. In regions of high LD, such a SNP could tag multiple causal variants located in different genes with completely different functions and thus lead to a spurious CP finding (FIG. 1f). This issue can be demonstrated by the major histocompatibility complex region that has been implicated in many complex traits, including autoimmune diseases[34]. This region contains more than 100 genes and has high levels of LD. A CP association in this region will probably tag multiple genes, and thus it can be particularly challenging to distinguish between biological and spurious pleiotropy.

## Analytical strategies to identify CP effects

Many methods have been proposed to test the association between a genetic variant and multiple phenotypes. These can be broadly classified into `multivariate analyses` and `univariate analyses`, and the most suitable approach depends on the circumstances. These methods facilitate the initial identification of CP effects, and details of study design considerations to minimize spurious associations are discussed in BOX 1. The subsequent approaches for classifying and characterizing the identified CP effects are discussed later.

Before searching for specific CP variants, it is possible first to implement a polygenic approach that uses all or a large proportion of SNPs genome-wide to establish genetic overlap between two traits. For example, common genetic variants were found to underlie schizophrenia and bipolar disorder (as shown by `polygenic scoring`)[28] and also type 2 diabetes and hypertension (as shown by a `linear mixed-effect model`)[41]. Note that both approaches assess whether pleiotropy exists between phenotypes but do not point to any particular variant or region of the genome.

### Multivariate approaches

Multivariate analyses jointly analyse more than one phenotype in a unified framework and test for the association of multiple phenotypes with a genetic variant. Because most multivariate methods require that all phenotypes be measured on the same individual, they

are only well suited for studies in which subjects are phenotyped across various diseases (for example, large cohort studies or cross-sectional studies). This is usually not feasible for diseases with a low prevalence, which are typically collected using a case–control study design. However, if phenotyping individuals on all traits is possible, this allows the investigation of the correlations between the traits themselves, rather than just testing of associations between genetic variants and the traits. One complication of multivariate methods is that they generally require pooling of individual-level data, and this may not be possible without reacquiring patient consent, implementing privacy protection measures and seeking additional ethical review board approval.

Numerous multivariate approaches have been proposed for testing the association between a genetic variant and multiple phenotypes, particularly for correlated phenotypes. The choice of method will largely depend on the types of traits (that is, continuous, categorical or binary) included in the analysis. For continuous phenotypes, a multivariate regression framework (such as a multivariate analysis of variance) can be used, but the approach requires that the phenotypes are approximately normally distributed. Several methods extend the regression framework, using variations of generalized estimating equations (GEE), to allow non-normally distributed phenotypes[42–44]. To model multiple categorical phenotypes (for example, multiple binary disease traits), a log-linear model[45] and a Bayesian network[46] have been used. In addition, there are several approaches that can accommodate a mixture of continuous and categorical phenotypes[44,47,48]. Ordinal regression[47] uses the genotype as the outcome variable and the set of phenotypes as the predictors. A non-parametric approach has been developed for a mixture of phenotypes but cannot incorporate additional predictors beyond the genetic variant[48].

Other approaches include a dimension reduction technique on the phenotypes before testing the association with the genetic variant. Principal components analysis (PCA)[49–51] extracts linear combinations (that is, principal components) of the traits that can be used as the phenotypes in a genetic association analysis. Canonical correlation analysis[52] extracts a linear combination of the phenotypes that explains the largest amount of covariation with the genetic variant. The weights for the linear combination will differ for each genetic variant, in contrast to PCA, and will provide information about which phenotypes are most strongly related.

These and other multivariate methods have recently been reviewed, and we refer the reader to those summaries[53] for further details.

## Univariate approaches

It is also possible to combine results from standard univariate analyses (such as GWAS associations between variants and single phenotypes) by combining these associations across various phenotypes to identify those variants that are associated with multiple phenotypes (summarized in TABLE 2). Thus, univariate approaches are well suited to analysing existing GWAS results, including the plethora of wellpowered GWASs conducted[54] by consortia already organizing themselves into cross-disease groups (such as the Psychiatric Genomics Consortium[7]). This will be especially important for rare diseases,

which are less likely to be ascertained in cohort studies. As the genetic effects for most complex traits are small[54], combining results across studies of different phenotypes can improve the power of detecting CP associations. This improvement in power will generally outweigh the advantages of using one study in which individuals are phenotyped on all traits. Another advantage of univariate approaches is that, unlike multivariate approaches, most of them are based on `summary statistics`, which do not divulge individual-level data and thus maintain participant confidentiality.

The simplest univariate approach is to take the known genome-wide-significant associations between variants and individual phenotypes and to compare the results across multiple phenotypes. CP effects are then declared at markers that satisfy the significance threshold for multiple traits. Alternatively, the set of genomewide-significant SNPs for one phenotype can be tested for association with other phenotypes; in this case, the significance level for multiple testing is adjusted only for the number of tested SNPs rather than for SNPs genome-wide. Both of these approaches require robust discovery as a starting point: because the known associations are probably only a subset of the true associations (even in traits for which large sample sizes have been analysed[28,55]), these analyses are fairly underpowered and will overlook SNPs that are only moderately associated across a set of phenotypes.

Variations on meta-analysis have also been adapted for CP effect detection. Traditional meta-analysis approaches combine evidence for association with the same phenotype across numerous studies; for discovering CP effects, the evidence for association is combined across studies of multiple phenotypes. Meta-analytical approaches aggregate summary statistics from individual studies into one statistic to test for CP effects and can be applied genome-wide or on a pre-specified set of SNPs. Broadly speaking, these methods can be split into two groups. First, those methods based on association *P* values ignore allelic effect direction (a positive versus negative effect on the trait) and `effect heterogeneity` (different effect sizes across traits). Second, methods based on the effect sizes are sensitive to allelic effect direction and effect size. We note that in GWASs in which effect sizes are generally very small, accounting for effect heterogeneity may be of less concern.

The simplest meta-analytical approach[56] aggregates *P* values across phenotypes in different studies to test the null hypothesis that the genetic variant is not associated with any phenotype. Note that this approach (which is similar to most methods in this section) does not explicitly test for CP effects, as a significant association could be driven by one phenotype as opposed to two or more phenotypes.

The cross-phenotype meta-analysis (CPMA) statistic[14] also uses association *P* values and tests whether the observed *P* values deviate from the expected distribution of *P* values under the null hypothesis of no additional associations beyond those already known. Because the alternative hypothesis includes only models in which two or more of the phenotypes are associated with the SNP, this approach explicitly tests for CP effects. It is also worth noting that this approach benefits from increased numbers of phenotypes, making it particularly well suited to broad phenotypic surveys.

Standard meta-analysis based on effect estimates is commonly used to combine evidence of association across multiple GWASs for the same phenotype[54,57,58] and has also been used to combine evidence across multiple phenotypes[7]. Fixed-effects meta-analysis assumes that the genetic variant has the same effect on each phenotype, whereas random-effects meta-analysis allows the genetic effect to differ across phenotypes. Although random-effects meta-analysis incorporates a moderate level of effect heterogeneity, it is not well suited for situations in which the genetic variant has opposite effects on different phenotypes. In addition, both will have lower power when only a subset of analysed phenotypes is associated.

The subset-based meta-analysis[59] extends standard fixed-effects meta-analysis to allow for opposite effects and to include situations in which association is only to a subset of traits. This method exhaustively evaluates all possible combinations of non-null models for association, selects the strongest association and then adjusts for the multiple comparisons generated by the search. At present, this is the only method that identifies which traits a variant influences (through the model selection step), but this advantage comes with a steep multiple testing price: the number of possible non-null combinations to be adjusted for exponentially increases with the number of traits selected so that detection power decreases for even moderate phenotype counts.

Several groups have proposed extensions to O'Brien's linear combination test[60], which uses a weighted sum of the univariate test statistics. The extensions[61,62] account for effect heterogeneity by allowing the weights to differ by phenotype and mainly differ in how they arrive at those weights. These approaches were specifically developed for correlated traits measured in the same individuals and simplify to standard meta-analysis if the underlying data are taken from independent studies[61]. Similarly to the O'Brien's test, the 'Trait-based Association Test that uses Extended Simes' (TATES) procedure[63] was developed to detect effects across correlated traits measured in the same individuals but in contrast uses the *P* value for each trait. For each variant, the approach takes the minimum *P* value across the set of univariate tests carried out on each phenotype and then applies a weight to the *P* value to account for the number phenotypes tested and their correlation.

The 'Pleiotropy Regional Identification Method' (PRIMe)[64] searches for regions of the genome that contain genetic variants associated with multiple traits but does not require that the same genetic variant be associated with multiple phenotypes. For each region, the approach calculates a pleiotropic index as the number of traits that have at least one SNP with a univariate *P* value less than $P_S$ (which is a pre-defined threshold) and then assesses the significance of the pleiotropic index. A related approach assesses whether `expression quantitative trait loci` (eQTLs) overlap disease associations; identifying effects on gene expression that result from variants in the identified region increases the confidence that this region harbours causal molecular candidates underlying the trait[65].

Overall, choosing the appropriate statistical approach for detecting a CP variant depends on study design, the type of phenotype, assumptions on effect heterogeneity and other factors that are summarized in TABLE 2. We will not enumerate all possible scenarios but aim to provide some general guidelines. When focusing on a small number of phenotypes (such as

five or less) that are of the same type (for example, all binary or all continuous), standard meta-analysis can be used, but this has the disadvantage that SNPs with opposite effects on the phenotypes will be missed. CPMA can accommodate opposite risk effects and different types of phenotypic traits and is well suited for moderate to large numbers of phenotypes (such as more than five). After conducting standard meta-analysis or CPMA genome-wide, a model selection technique (for example, subset-based meta-analysis) can be applied to the top selection of SNPs to refine the association and to identify which of the phenotypes is driving the signal (BOX 2). When there are overlapping subjects (for example, shared controls) across studies, the overlapping subjects can be split across the different studies, and then univariate tests are carried out so that each subject is used only once. Then the tests can be assumed to be uncorrelated. Alternatively, Lin *et al*.[66] have provided an adjustment for overlapping subjects for standard meta-analysis, and Bhattacharjee *et al*.[59] have proposed a similar extension to the subset-based metaanalysis. Finally, if the phentoypes are measured on the same subjects, alternative methods can be used, including the extensions to the O'Brien linear combination test, the TATES procedure or one of the many multivariate approaches.

## Distinguishing and characterizing CP effects

The forms of pleiotropy are important to distinguish because they imply distinct molecular mechanisms and have different implications for disease risk and pathogenesis. Strategies to achieve this are described below, and further functional characterization of CP-effect loci is discussed in BOX 3.

### Fine mapping to distinguish biological and spurious pleiotropy

Careful study design is required in order to minimize the identification of spurious pleiotropy caused by artefactual CP associations (BOX 1); additionally, when feasible, `fine mapping` of the region that surrounds a CP effect can help to discriminate spurious from biological pleiotropy. Such mapping is used to locate the causal variant or variants that are responsible for a CP effect. If a single variant or variants in the same gene are causal for the diseases, this indicates biological pleiotropy (FIG. 1a–c), whereas causal variants in different genes that are in LD is suggestive of spurious pleiotropy (FIG. 1f). Fine mapping can also aid in distinguishing the different forms of biological pleiotropy and, in particular, can identify whether the observed CP association is driven by one variant (FIG. 1a) or multiple variants (FIG. 1b,c) in the same gene that is associated with different phenotypes. This can be particularly challenging when two variants in the same gene are in strong LD and may be related to different diseases (FIG. 1b), because these variants will typically co-occur in individuals, such that the effects of each individual SNP will rarely be able to be dissected. For common diseases that can co-occur in the same individual, variants for the first disease can be mapped in the presence of the second disease and then in its absence to establish which variant is related to the first disease (and vice versa).

Custom genotyping arrays have been designed to fine-map regions identified in GWASs for immunemediated traits (Immunochip[67]) and for metabolic, cardiovascular and anthropometric traits (Metabochip[68]). This provides a low-cost alternative to sequencing and allows for fine mapping in large sample sizes.

Finally, it is worth noting that in many cases, establishing whether a variant is truly causal cannot be established by fine mapping alone and requires biological and animal studies to determine the exact function of the variant (BOX 3).

### Identifying mediated pleiotropy

In cases of potential mediated pleiotropy, the association between the genetic variant and the second phenotype (that is, target phenotype) can be tested while adjusting or stratifying by the first (that is, intermediate phenotype). If the association persists (that is, if the variant is associated with the target phenotype even when the intermediate phenotype is not present), then the CP effect is probably not fully mediated. However, this approach can produce biased results when the phenotypes share a `confounding factor` that is influenced by the genetic variant[69]. To address this shortcoming, approaches using causal inference methodology have been developed to test whether a genetic variant influences the target phenotype through a path that does not involve the intermediate phenotype[69–71]. Such an approach demonstrates that the association between SNPs at 15q25.1 with both smoking and lung cancer mostly reflects direct effects on each phenotype, rather than mediated pleiotropy[72].

More generally, identifying mediated pleiotropic genetic effects can provide a tool by which causation and correlation can be teased apart under some conditions in an approach called Mendelian randomization[37,73,74] (BOX 4). This framework for causal inference tests whether the intermediate phenotype causally affects the target phenotype. Specifically, if the effect of a genetic variant can be taken as a proxy for the intermediate phenotype, this is used to establish the causal relationship between the intermediate phenotype and the disease. Using Mendelian randomization, Voight and colleagues[37] found that LDL levels causally affect myocardial infarction risk, whereas high-density lipoprotein (HDL) levels do not. This counter-intuitive result suggests that low HDL may be a consequence rather than a cause of myocardial infarction risk, thus challenging the established view that increasing the levels of HDL cholesterol will uniformly lower the risk of myocardial infarction. However, we note that the assumptions underlying Mendelian randomization are quite strong (BOX 4), and thus extreme care needs to be taken in experimental design and data interpretation.

## Clinical implications of CP effects

Characterizing the molecular mechanisms of CP effects (BOX 3) will undoubtedly expand our understanding of the underlying biology of complex diseases and will have clinical implications for drug discovery. First, characterizing CP effects may have clinically relevant implications for the classification (nosology) of medical disorders. For example, psychiatric disorders are currently defined as distinct syndromes on the basis of their constellations of signs and symptoms. As noted earlier, however, recent GWASs[7,28] have demonstrated shared heritability among many of these disorders[29,75,76]. As further studies provide a more comprehensive account of the distinct and overlapping `genetic architecture` of psychiatric disorders, the goal of an aetiology-based classification may become more feasible. Of note, imperfect nosology poses a challenge for teasing apart biological pleiotropy from spurious pleiotropy (particularly the bias resulting from misdiagnosis) as the

distinction between two disorders may not be aetiologically valid. In such cases, the pleiotropy may be real, but the diagnostic categories are in fact spurious.

The growing catalogue of genetic variants with pleiotropic effects has important implications for genetic testing and personal genomics. As genetic information is increasingly integrated into medical practice, clinicians and medical genetics professionals will need to be aware that genetic tests for one disease may have implications for risks of other diseases. In some cases, discovery of these secondary risks may emerge well after the original test information has been provided, thus complicating the process of genetic counselling and raising complex ethical and 'duty to warn' issues. At the same time, the growth of direct-to-consumer genetic tests will mean that an increasing number of individuals will be confronted with CP risk information without the benefit of genetic counselling. The case of *APOE* provides a familiar example of a common variant with well-established CP effects. The *APOE4* allele is a known risk factor for both atherosclerotic heart disease and Alzheimer's disease but has also been shown to exert a protective effect on risk of age-related macular degeneration[77]. Very little research is available to evaluate the psychological impact of such 'competing risk' information. In addition, accurately characterizing CP effects and distinguishing between biological and mediated pleiotropy will affect how this information is interpreted and used in clinical practice. For example, if a patient carries a variant that directly affects myocardial infarction through LDL, the mediated relationship provides clinicians with a more proximal target for the prevention of myocardial infarction. Furthermore, distinguishing between CP effects caused by single versus multiple variants can improve the accuracy of these genetic tests and the interpretation of results. For example, although the same gene may be implicated in multiple diseases, if distinct variants in that gene are differentially associated with alternative diseases, then testing for both variants might provide separate risk information for each disease.

In the realm of therapeutics, the existence of common pathological mechanisms in distinct disorders may suggest new opportunities and challenges for drug development. Drugs developed for one disorder could be repurposed to treat another disorder if the therapeutic target is found to be common to the biology of both disorders. In such cases, a gene or multiple genes in a pathway might be considered to be pleiotropic if they affect more than one phenotype, regardless of whether the specific variants are shown to have CP effects. For example, the finding that the L-type calcium channel subunit gene *CACNA1C* is a risk gene for bipolar disorder[78] has revived interest in trials of calcium channel antagonist antihypertensive drugs as possible mood disorder treatments (R. H. Perlis, personal communication). Alternatively, a drug targeting a shared pathway could be beneficial for one disease and detrimental for others; this scenario could result in 'off-target' effects at the disease level despite being on-target at the pharmacological level. For example, several genes have opposing effects on autoimmune disorders[79–81], suggesting that drugs modulating these gene products to treat one disorder could have unintended adverse effects on another. This is exemplified by the utility of treatments targeted to tumour necrosis factor (TNF) in Crohn's disease and rheumatoid arthritis but their counter-indication in multiple sclerosis. The adverse effect on multiple sclerosis is also supported by evidence of a genetic

variant identified in GWASs that increases the risk of disease and mimics the effect of TNF-targeted treatments[82].

## Conclusions and future directions

An exciting picture is emerging of startling genetic overlap between seemingly unrelated diseases and traits. The promise is twofold: using ever-larger sample sizes across genetic cohorts will further increase discoveries of genetic association, and the patterns of sharing will help to sort associations into discrete pathways, which will further our understanding of biology and disease. In this Review, we have outlined analytical strategies to discover CP effects systematically in existing GWAS data sets as the first step in this direction. Several advances will be instrumental in allowing us to reap the full benefits of shared genetic architecture across traits: analytical frameworks, such as those we describe, must be developed, tested and implemented; multi-disease mega-consortia must be formed to pool data across traits; and systems-level approaches must be developed to characterize the molecular mechanisms perturbed by common CP associations of modest effect (BOX 3).

This Review has focused on the detection of CP effects, but functionally characterizing identified variants and understanding the underlying mechanism remains a major challenge in the field (BOX 3). Although many resources are available for characterizing protein-coding variants, experiments in animal or cellular models are generally necessary to establish causality. The Encyclopedia of DNA Elements (ENCODE) project provides a valuable resource for characterizing nonprotein-coding variants and regulatory elements and has found that most GWAS associations overlap a functional region[25]. In addition, examining eQTLs in the relevant tissue for each phenotype of a CP effect can help to elucidate the functional consequence and to distinguish between biological and spurious pleiotropy. Finally, network- based approaches[83,84] have highlighted the importance of pleiotropy in human disease, and understanding CP effects in the context of these models can provide insight into the mechanisms of shared pathophysiology. For example, proteins involved in the same disease are more likely to interact with each other[83], pathophenotypes within the same disease class are more likely to share genes[84], and increased comorbidity has been identified among diseases that are metabolically linked[85].

As the field moves towards sequencing-based association studies, we will have the opportunity directly to identify the causal alleles underlying the CP effects identified in GWASs and thus to distinguish between the different types of pleiotropy more accurately. The current focus on whole-exome sequencing will probably bias findings towards gene-centric pleiotropic effects, whereas whole-genome sequencing will provide a more robust survey of the genomic landscape for CP effects. Sequencing will also allow us to interrogate lower-frequency variants (which are typically not represented on SNP-genotyping microarrays) for CP effects, and some of these variants are likely to have higher penetrance than those found in GWASs. In addition, the observed comorbidity between mapped Mendelian disorders and complex traits can be exploited by carrying out focused sequencing of the mapped region. For example, comorbidity between Gaucher's disease and Parkinsonism led to the identification of risk alleles for Parkinson's disease in *GBA*, which is the gene implicated in Gaucher's disease[86].

Extending observations of CP effects to a wider range of phenotypes is an emerging area. Systematic and unbiased phenome-wide association studies (PheWASs) are now beginning in which a SNP with an established association with a phenotype is tested for association with hundreds of other phenotypes[87,88]. The Population Architecture using Genomics and Epidemiology (PAGE) network[89] is a large-scale collaboration for harmonizing phenotypes across eight epidemiological studies and five ethnic groups for the purpose of conducting PheWASs on replicated GWAS hits[90]. Other efforts aim to analyse a broad range of phenotypes that are extracted from electronic medical records[88,91]. These approaches will increase our understanding of the extent of shared genetics among traits and our global understanding of phenotypes as a range of inter-related manifestations of biological mechanisms rather than isolated events.

## Acknowledgments

This work was supported in part by the US National Institute of Mental Health (NIMH) grants R01-MH079799 and K24MH094614 (both to J.W.S.).

## Glossary

| | |
|---|---|
| **Genome-wide association studies** | (GWASs). Studies in which hundreds of thousands (or millions) of genetic markers are tested for association with a phenotypic trait; they are an unbiased approach to survey the entire genome for disease-associated regions using common variation. |
| **Genome-wide-significant** | A term describing the statistical significance threshold that accounts for multiple testing in GWASs. |
| **Complex traits** | Traits controlled by a combination of many genes and environmental factors. |
| **Pleiotropy** | A gene or genetic variant that affects more than one phenotypic trait. |
| **Heritability** | The proportion of phenotypic variance attributed to genetic differences among individuals in a population. |
| **Colocalizing** | Different genetic variants in high linkage disequilibrium located in the same gene that affect different phenotypes. |
| **Single-nucleotide polymorphisms** | Single-nucleotides in the genome that vary across individuals in the population. |
| **Linkage disequilibrium** | (LD). The correlation between genetic markers owing to limited recombination. |
| **Copy number variants** | Regions of the genome in which the copy number is polymorphic (for example, deletions and duplications) across individuals. |
| **Polygenic** | Controlled by many genes. |

| | |
|---|---|
| **Population stratification** | A source of bias in genome-wide association studies that occurs when a phenotype and the allele frequency of a single-nucleotide polymorphism vary owing to ancestral differences. |
| **Batch effect** | Systematic biases in the data that arise from differences in sample handling. |
| **Genotype imputation** | Inference of missing genotypes or untyped single-nucleotide polymorphisms using statistical techniques. |
| **Ascertainment bias** | A consequence of collecting a nonrandom subsample with a systematic bias so that results based on the subsample are not representative of the entire sample. |
| **Tag SNPs** | Single-nucleotide polymorphisms (SNPs) chosen to represent a region of the genome owing to strong linkage disequilibrium. |
| **Multivariate analyses** | The simultaneous inclusion of two or more phenotypes in one analysis when testing the association with a genetic variant. |
| **Univariate analyses** | Tests of association between one phenotype and a genetic variant. |
| **Polygenic scoring** | A score that aggregates the number of risk alleles a subject carries weighted by the effect size of the allele for a particular trait. The risk allele and effect size for each single-nucleotide polymorphism is generally taken from a genome-wide association study of an independent study. |
| **Linear mixed-effect model** | A linear model that contains both fixed and random effects. This type of model can be used to estimate genetic correlation between traits using a genome-wide set of single-nucleotide polymorphisms. |
| **Cohort studies** | Observational studies in which defined groups of people (the cohorts) are followed over time and outcomes are compared in subsets of the cohort who were exposed to different levels of factors of interest. These studies can either be prospectively or retrospectively carried out from historical records. |
| **Cross-sectional studies** | Studies in which data are collected on subjects at one specific point in time and subjects are not selected for a particular trait or exposure. |
| **Case–control study** | Compares cases (that is, a selected group of individuals: for example, those diagnosed with a disorder) with controls (that is, a comparison group of individuals: for example, those who are not diagnosed with the disorder). Genome-wide association case–control studies test whether genetic marker allele frequencies differ between cases and controls. |

| | |
|---|---|
| **Generalized estimating equations** | A statistical technique used to estimate regression parameters that does not require the joint distribution of the variables to be fully specified. |
| **Log-linear model** | A statistical model that captures the dependence among a set of categorical variables. |
| **Bayesian network** | A network that captures relationships between variables or nodes of interest (for example, phenotypes and SNPs). Bayesian networks can incorporate prior information in establishing relationships between variables. |
| **Ordinal regression** | A regression model in which the outcome variable is ordinal. |
| **Non-parametric approach** | A statistical analysis method that does not rely on specific distributional assumptions (for example, normality) for the variables being analysed. |
| **Principal components analysis** | A statistical method used to simplify data sets by transforming a series of correlated variables into a smaller number of uncorrelated factors. It is also commonly used to infer continuous axes of variation in genetic data, often representing genetic ancestry. |
| **Summary statistics** | A statistic that summarizes a set of observations. In the context of genome-wide association studies, meta-analyses can be carried out solely by using summary statistics and typically include estimates of the effect size (for example, odds ratio) and standard error. |
| **Effect heterogeneity** | Different effect sizes across phenotypes. |
| **Expression quantitative trait loci** | Loci at which genetic allelic variation is associated with variation in gene expression. |
| **Fine mapping** | Extensively genotyping or sequencing a region of the genome that was identified in genome-wide association studies to identify the causal variant. |
| **Confounding factor** | A variable (for example, batch effects or population structure) that is associated with both the genotype and the phenotype of interest and can give rise to a spurious association. |
| **Genetic architecture** | A genetic model (that is, the number of single-nucleotide polymorphisms, effect sizes, allele frequency, and so on) underlying a phenotypic trait. |

## References

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl Acad. Sci. USA. 2009; 106:9362–9367. [PubMed:

19474294] Characteristics of reported GWAS results listed in the US National Human Genome Research Institute (NHGRI) catalogue are discussed in this paper.

2. Plenge RM, et al. Replication of putative candidategene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. Am. J. Hum. Genet. 2005; 77:1044–1060. [PubMed: 16380915]
3. Barrett JC, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genet. 2008; 40:955–962. [PubMed: 18587394]
4. Kyogoku C, et al. Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. Am. J. Hum. Genet. 2004; 75:504–507. [PubMed: 15273934]
5. Todd JA, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nature Genet. 2007; 39:857–864. [PubMed: 17554260]
6. Fletcher O, Houlston RS. Architecture of inherited susceptibility to common cancer. Nature Rev. Cancer. 2010; 10:353–361. [PubMed: 20414203]
7. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013; 381:1371–1379. [PubMed: 23453885] This paper presents a genome-wide analysis of CP associations across five psychiatric disorders.
8. Stearns FW. One hundred years of pleiotropy: a retrospective. Genetics. 2010; 186:767–773. [PubMed: 21062962] This is a historical review of pleiotropy.
9. Wagner GP, Zhang J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nature Rev. Genet. 2011; 12:204–213. [PubMed: 21331091] This excellent Review discusses pleiotropy in model organisms and the implications for evolution.
10. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. Major depression and generalized anxiety disorder. Same genes, (partly) different environments? Arch. Gen. Psychiatry. 1992; 49:716–722. [PubMed: 1514877]
11. Criswell LA, et al. Analysis of families in the Multiple Autoimmune Disease Genetics Consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. Am. J. Hum. Genet. 2005; 76:561–571. [PubMed: 15719322]
12. Eaton WW, Rose NR, Kalaydjian A, Pedersen MG, Mortensen PB. Epidemiology of autoimmune diseases in Denmark. J. Autoimmun. 2007; 29:1–9. [PubMed: 17582741]
13. Sivakumaran S, et al. Abundant pleiotropy in human complex diseases and traits. Am. J. Hum. Genet. 2011; 89:607–618. [PubMed: 22077970]
14. Cotsapas C, et al. Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 2011; 7:e1002254. [PubMed: 21852963] Systematic evaluation of CP associations is carried out in this study across seven autoimmune diseases and application of CPMA method.
15. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ. Autoimmune disease classification by inverse association with SNP alleles. PLoS Genet. 2009; 5:e1000792. [PubMed: 20041220]
16. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233] This is the largest study of Crohn's disease and ulcerative colitis and identifies more than 100 CP associations.
17. Thorleifsson G, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nature Genet. 2009; 41:18–24. [PubMed: 19079260]
18. Iles MM, et al. A variant in FTO shows association with melanoma risk not due to BMI. Nature Genet. 2013; 45:428–432. [PubMed: 23455637]
19. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature Genet. 2011; 43:333–338. [PubMed: 21378990]
20. The Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nature Genet. 2011; 43:339–344. [PubMed: 21378988]
21. Shete S, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nature Genet. 2009; 41:899–904. [PubMed: 19578367]
22. Yasuno K, et al. Genome-wide association study of intracranial aneurysm identifies three new risk loci. Nature Genet. 2010; 42:420–425. [PubMed: 20364137]

23. Tomlinson I, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nature Genet. 2007; 39:984–988. [PubMed: 17618284]
24. Thomas G, et al. Multiple loci identified in a genomewide association study of prostate cancer. Nature Genet. 2008; 40:310–315. [PubMed: 18264096]
25. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012; 22:1748–1759. [PubMed: 22955986]
26. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012; 148:1223–1241. [PubMed: 22424231]
27. Heinzen EL, et al. Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes. Am. J. Hum. Genet. 2010; 86:707–718. [PubMed: 20398883]
28. Purcell SM, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]
29. Lichtenstein P, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet. 2009; 373:234–239. [PubMed: 19150704]
30. Rieck M, et al. Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. J. Immunol. 2007; 179:4704–4710. [PubMed: 17878369]
31. Menard L, et al. The *PTPN22* allele encoding an R620W variant interferes with the removal of developing autoreactive B cells in humans. J. Clin. Invest. 2011; 121:3635–3644. [PubMed: 21804190]
32. Zhang J, et al. The autoimmune disease-associated PTPN22 variant promotes calpain-mediated Lyp/Pep degradation associated with lymphocyte and dendritic cell hyperresponsiveness. Nature Genet. 2011; 43:902–907. [PubMed: 21841778]
33. Behrens TW. Lyp breakdown and autoimmunity. Nature Genet. 2011; 43:821–822. [PubMed: 21874033]
34. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature Rev. Genet. 2009; 10:43–55. [PubMed: 19092835]
35. Pomerantz MM, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature Genet. 2009; 41:882–884. [PubMed: 19561607]
36. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential *in vivo* activity to a MYC enhancer. Genome Res. 2010; 20:1191–1197. [PubMed: 20627891]
37. Voight BF, et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. Lancet. 2012; 380:572–580. [PubMed: 22607825] This paper presents an example of Mendelian randomization using results from GWASs.
38. Hung RJ, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008; 452:633–637. [PubMed: 18385738]
39. Thorgeirsson TE, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008; 452:638–642. [PubMed: 18385739]
40. Chanock SJ, Hunter DJ. Genomics: when the smoke clears. Nature. 2008; 452:537–538. [PubMed: 18385720]
41. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012; 28:2540–2542. [PubMed: 22843982]
42. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42:121–130. [PubMed: 3719049]
43. Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics. 2003; 4:195–206. [PubMed: 12925516]
44. Liu J, Pei Y, Papasian CJ, Deng HW. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet. Epidemiol. 2009; 33:217–227. [PubMed: 18924135]

45. Lee PH, et al. Modifiers and subtype-specific analyses in whole-genome association studies: a likelihood framework. Hum. Hered. 2011; 72:10–20. [PubMed: 21849790]
46. Hartley SW, Monti S, Liu CT, Steinberg MH, Sebastiani P. Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. Front. Genet. 2012; 3:176. [PubMed: 22973300]
47. O'Reilly PF, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE. 2012; 7:e34861. [PubMed: 22567092]
48. Zhang H, Liu CT, Wang X. An association test for multiple traits based on the generalized Kendall's tau. J. Am. Stat. Assoc. 2010; 105:473–481. [PubMed: 20711441]
49. Ott J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. Hum. Hered. 1999; 49:106–111. [PubMed: 10077732]
50. Lange C, et al. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat. Appl. Genet. Mol. Biol. 2004; 3 Article 17.
51. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet. Epidemiol. 2008; 32:9–19. [PubMed: 17922480]
52. Ferreira MA, Purcell SM. A multivariate test of association. Bioinformatics. 2009; 25:132–133. [PubMed: 19019849]
53. Shriner D. Moving toward system genetics through multiple trait analysis in genome-wide association studies. Front. Genet. 2012; 3:1. [PubMed: 22303408] This is a review of multivariate approaches for detecting CP associations.
54. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nature Rev. Genet. 2009; 10:318–329. [PubMed: 19373277]
55. Stahl EA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature Genet. 2012; 44:483–489. [PubMed: 22446960]
56. Fisher, RA. Statistical Methods for Research Workers. Oliver & Boyd; 1925.
57. Kavvoura FK, Ioannidis JP. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. Hum. Genet. 2008; 123:1–14. [PubMed: 18026754]
58. de Bakker PI, et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. 2008; 17:R122–R128. [PubMed: 18852200]
59. Bhattacharjee S, et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am. J. Hum. Genet. 2012; 90:821–835. [PubMed: 22560090]
60. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984; 40:1079–1087. [PubMed: 6534410]
61. Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. Biostatistics. 2003; 4:223–229. [PubMed: 12925518]
62. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet. Epidemiol. 2010; 34:444–454. [PubMed: 20583287]
63. van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013; 9:e1003235. [PubMed: 23359524]
64. Huang J, Johnson AD, O'Donnell CJ. PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. Bioinformatics. 2011; 27:1201–1206. [PubMed: 21398673]
65. Nica AC, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010; 6:e1000895. [PubMed: 20369022]
66. Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. Am. J. Hum. Genet. 2009; 85:862–872. [PubMed: 20004761]
67. Cortes A, Brown MA. Promise and pitfalls of the immunochip. Arthritis Res. Ther. 2011; 13:101. [PubMed: 21345260]

68. Voight BF, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012; 8:e1002793. [PubMed: 22876189]
69. Vansteelandt S, et al. On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. Genet. Epidemiol. 2009; 33:394–405. [PubMed: 19219893]
70. Lipman PJ, Lange C. CGene: an R package for implementation of causal genetic analyses. Eur. J. Hum. Genet. 2011; 19:1292–1294. [PubMed: 21731061]
71. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am. J. Epidemiol. 2010; 172:1339–1348. [PubMed: 21036955]
72. VanderWeele TJ, et al. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. Am. J. Epidemiol. 2012; 175:1013–1020. [PubMed: 22306564]
73. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med. 2008; 27:1133–1163. [PubMed: 17886233]
74. Glymour MM, Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. Am. J. Epidemiol. 2012; 175:332–339. [PubMed: 22247045]
75. McGuffin P, et al. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. Arch. Gen. Psychiatry. 2003; 60:497–502. [PubMed: 12742871]
76. Rommelse NN, Franke B, Geurts HM, Hartman CA, Buitelaar JK. Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder. Eur. Child Adolesc. Psychiatry. 2010; 19:281–295. [PubMed: 20148275]
77. McKay GJ, et al. Evidence of association of APOE with age-related macular degeneration: a pooled analysis of 15 studies. Hum. Mutat. 2011; 32:1407–1416. [PubMed: 21882290]
78. Ferreira MA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nature Genet. 2008; 40:1056–1058. [PubMed: 18711365]
79. Wang K, et al. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. Hum. Mol. Genet. 2010; 19:2059–2067. [PubMed: 20176734]
80. Smyth DJ, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med. 2008; 359:2767–2777. [PubMed: 19073967]
81. Zhernakova A, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 2011; 7:e1002004. [PubMed: 21383967]
82. Gregory AP, et al. TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature. 2012; 488:508–511. [PubMed: 22801493]
83. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Rev. Genet. 2011; 12:56–68. [PubMed: 21164525]
84. Goh KI, et al. The human disease network. Proc. Natl Acad. Sci. USA. 2007; 104:8685–8690. [PubMed: 17502601] A first step is taken in this study towards the construction of the genotype–phenotype map in humans using known disease genes reported in OMIM (Online Mendelian Inheritance in Man)
85. Lee DS, et al. The implications of human metabolic network topology for disease comorbidity. Proc. Natl Acad. Sci. USA. 2008; 105:9880–9885. [PubMed: 18599447]
86. DePaolo J, Goker-Alpan O, Samaddar T, Lopez G, Sidransky E. The association between mutations in the lysosomal protein glucocerebrosidase and parkinsonism. Mov. Disord. 2009; 24:1571–1578. [PubMed: 19425057]
87. Denny JC, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover genedisease associations. Bioinformatics. 2010; 26:1205–1210. [PubMed: 20335276]
88. Denny JC, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am. J. Hum. Genet. 2011; 89:529–542. [PubMed: 21981779]

89. Pendergrass SA, et al. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. Genet. Epidemiol. 2011; 35:410–422. [PubMed: 21594894]
90. Pendergrass SA, et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the population architecture using genomics and epidemiology (PAGE) network. PLoS Genet. 2013; 9:e1003087. [PubMed: 23382687]
91. Rasmussen-Torvik LJ, et al. High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. Clin. Transl. Sci. 2012; 5:394–399. [PubMed: 23067351]
92. Smoller JW, Lunetta KL, Robins J. Implications of comorbidity and ascertainment bias for identifying disease genes. Am. J. Med. Genet. 2000; 96:817–822. [PubMed: 11121189]
93. Berkson J. Limitations of the application of fourfold table analysis to hospital data. Biometrics. 1946; 2:47–53. [PubMed: 21001024]
94. Wray NR, Lee SH, Kendler KS. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. Eur. J. Hum. Genet. 2012; 20:668–674. [PubMed: 22258521]
95. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev. Genet. 2008; 9:356–369. [PubMed: 18398418] This Review presents an overview of key considerations and challenges in GWASs.
96. Laurie CC, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. 2010; 34:591–602. [PubMed: 20718045]
97. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 2006; 38:904–909. [PubMed: 16862161]
98. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nature Rev. Genet. 2010; 11:459–463. [PubMed: 20548291]
99. Rosenberg NA, et al. Genome-wide association studies in diverse populations. Nature Rev. Genet. 2010; 11:356–366. [PubMed: 20395969]
100. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Rev. Genet. 2010; 11:499–511. [PubMed: 20517342]
101. Kann M. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Brief. Bioinform. 2010; 11:96–110. [PubMed: 20007728]
102. Adzhubei I, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7:248–249. [PubMed: 20354512]
103. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protoc. 2009; 4:1073–1081. [PubMed: 19561590]
104. Freedman M, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nature Genet. 2011; 43:513–518. [PubMed: 21614091]
105. Fehrmann R, et al. *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011; 7:e1002197. [PubMed: 21829388]
106. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 2011; 27:72–79. [PubMed: 21122937]
107. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008; 24:408–415. [PubMed: 18597885]
108. Baker M. Biorepositories: building better biobanks. Nature. 2012; 486:141–146. [PubMed: 22678297]
109. Cantor R, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am. J. Hum. Genet. 2010; 86:6–22. [PubMed: 20074509]
110. Eleftherohorinou H, et al. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PLoS Genet. 2009; 4:e8068.
111. Evans MJ, Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. Nature. 1981; 292:154–156. [PubMed: 7242681]

112. Smithies O, Gregg RG, Boggs SS, Koralewski MA, Kucherlapati RS. Insertion of DNA sequences into the human chromosomal β-globin locus by homologous recombination. Nature. 1985; 317:230–234. [PubMed: 2995814]
113. Thomas KR, Folger KR, Capecchi MR. High frequency targeting of genes to specific sites in the mammalian genome. Cell. 1986; 44:419–428. [PubMed: 3002636]
114. Li H, et al. *In vivo* genome editing restores haemostasis in a mouse model of haemophilia. Nature. 2011; 475:217–221. [PubMed: 21706032]
115. Esvelt KM, Wang HH. Genome-scale engineering for systems and synthetic biology. Mol. Syst. Biol. 2013; 9:641. [PubMed: 23340847]
116. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Rev. Genet. 2011; 12:628–640. [PubMed: 21850043]
117. Hu X, et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet. 2011; 89:496–506. [PubMed: 21963258]
118. Henderson NC, et al. Galectin-3 regulates myofibroblast activation and hepatic fibrosis. Proc. Natl Acad. Sci. USA. 2006; 103:5060–5065. [PubMed: 16549783]
119. Radosavljevic G, et al. The roles of galectin-3 in autoimmunity and tumor progression. Immunol. Res. 2012; 52:100–110. [PubMed: 22418727]
120. Honjo Y, Nangia-Makker P, Inohara H, Raz A. Down-regulation of galectin-3 suppresses tumorigenicity of human breast carcinoma cells. Clin. Cancer Res. 2001; 7:661–668. [PubMed: 11297262]
121. Shekhar MP, Nangia-Makker P, Tait L, Miller F, Raz A. Alterations in galectin-3 expression and distribution correlate with breast cancer progression: functional analysis of galectin-3 in breast epithelialendothelial interactions. Am. J. Pathol. 2004; 165:1931–1941. [PubMed: 15579437]
122. Baptiste TA, James A, Saria M, Ochieng J. Mechano-transduction mediated secretion and uptake of galectin-3 in breast carcinoma cells: implications in the extracellular functions of the lectin. Exp. Cell Res. 2007; 313:652–664. [PubMed: 17184769]
123. Nangia-Makker P, et al. Cleavage of galectin-3 by matrix metalloproteases induces angiogenesis in breast cancer. Int. J. Cancer. 2010; 127:2530–2541. [PubMed: 20162566]
124. Palmer TM, et al. Using multiple genetic variants as instrumental variables for modifiable risk factors. Stat. Methods Med. Res. 2012; 21:223–242. [PubMed: 21216802]
125. Duerr RH, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006; 314:1461–1463. [PubMed: 17068223]
126. Evans DM, et al. Interaction between ERAP1 and HLA B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA B27 in disease susceptibility. Nature Genet. 2011; 43:761–767. [PubMed: 21743469]
127. Silverberg MS, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genomewide association study. Nature Genet. 2009; 41:216–220. [PubMed: 19122664]
128. Strange A, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA C and ERAP1. Nature Genet. 2010; 42:985–990. [PubMed: 20953190]
129. Franke A, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genet. 2010; 42:1118–1125. [PubMed: 21102463]
130. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
131. Broderick P, et al. Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. Nature Genet. 2012; 44:58–61. [PubMed: 22120009]
132. Sanders SJ, et al. Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011; 70:863–885. [PubMed: 21658581]
133. Pober BR. Williams-Beuren syndrome. N. Engl. J. Med. 2010; 362:239–252. [PubMed: 20089974]

## Box 1 | Spurious CP associations and study design considerations

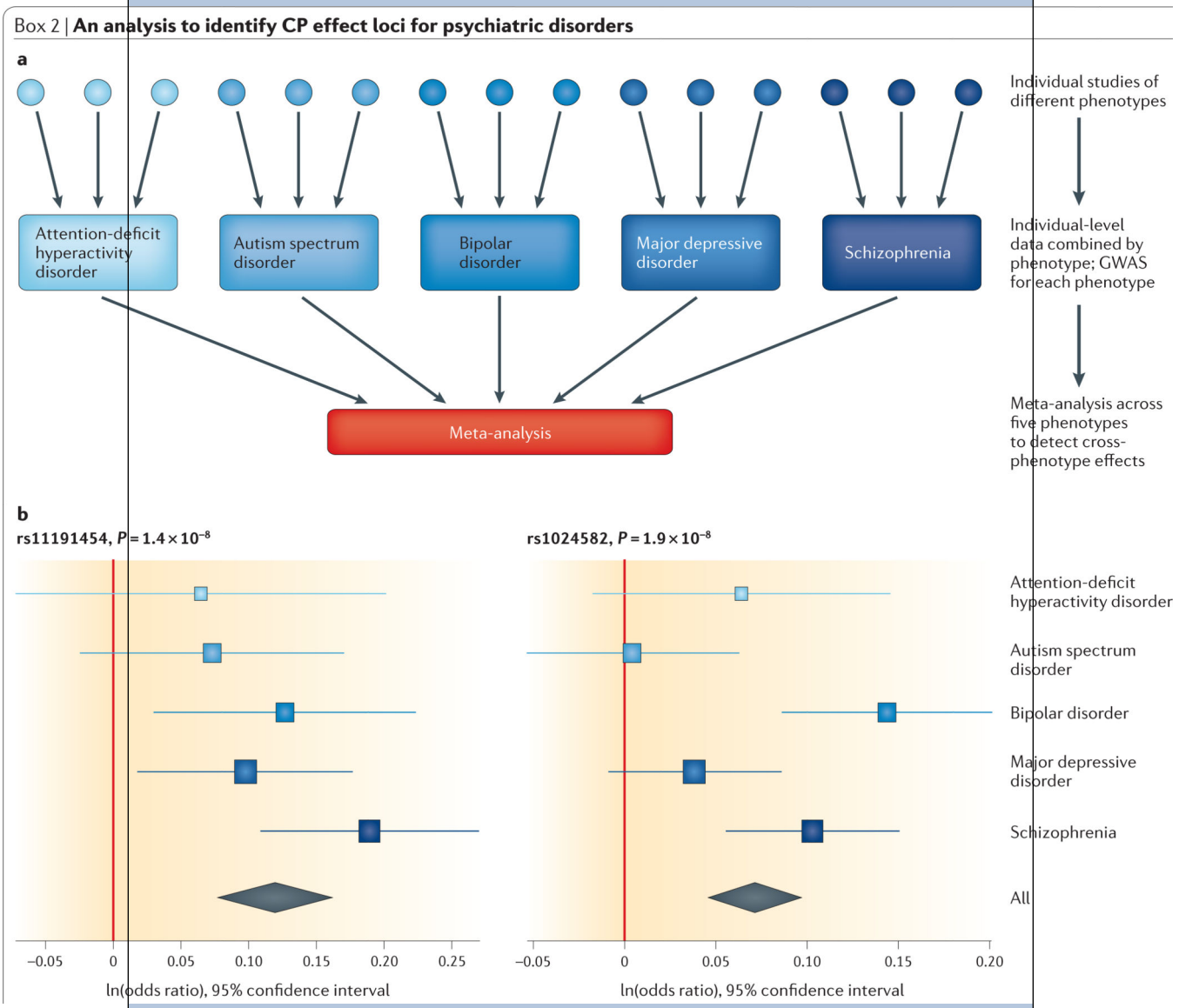
Ascertainment bias can induce spurious cross-phenotype (CP) effects and occurs when the recruitment of individuals with one phenotype increases the prevalence of a second unrelated phenotype[92] in the cohort and thus the spurious correlation between them (FIG. 1e). Such ascertainment bias (for example, Berkson's bias[93]) is common in clinically ascertained samples, as patients suffering from two conditions are often more likely to seek treatment than those suffering from one[92]. In addition, unaffected control individuals are often shared across multiple studies, and a spurious CP association could occur if an artefact (such as `population stratification` or `batch effects`) systematically biases the shared controls and not the cases.

Spurious CP effects between two phenotypes can also occur when subjects with one phenotype are systematically misclassified with a different phenotype (FIG. 1e). For example, patients with schizophrenia can sometimes be misdiagnosed with bipolar disorder and vice versa, and this could result in artefactual genetic correlation between the traits and hence could generate spurious CP effects. However, the misclassification rate must be quite high[94] to have a substantial impact on the genetic correlation. In the cases of schizophrenia and bipolar disorder, the misclassification rate would need to be larger than 20% to generate the genetic correlation (0.60) observed between the traits if the true genetic correlation were zero[94]. Although misclassification must be carefully considered as a source of bias during the design of the study, it is unlikely, in our opinion, that a substantial number of reported CP effects are caused by this type of bias.

These sources of bias are relevant for both multivariate and univariate analytical approaches (see the main text) and can be avoided with careful study design. General guidelines for study design of genome-wide association studies (GWASs) should be followed, including appropriate control selection, adequate quality control and adjustment for population stratification[95,96]. Population stratification is a major source of confounding in GWASs, and established methods of population stratification adjustment should be applied within studies[97,98]. In addition, appropriate adjustments for multiple testing[95] should be implemented to avoid false-positive CP associations, which could inflate the observed genetic overlap between traits. When testing for CP effects across different studies (which is particularly relevant for univariate approaches), similar populations should be used, as the underlying linkage disequilibrium structure varies across populations and the single-nucleotide polymorphism (SNP) of interest may differentially tag the causal variant[99]. Additionally, as the same individuals (particularly controls) might be used in multiple studies, the overlap across samples should be minimized and appropriately accounted for in the analysis. If the participant overlap is not accounted for, the estimates of effects can be biased, and the discovery power may be reduced[66]. A further consideration is that when combining phenotypes across studies, participants may have different sets of SNPs available owing to differences in genotyping arrays; in this setting, `genotype imputation`[100] can be used to obtain the same set of SNPs for all individuals. Special considerations are needed when the proportion of cases

differs across genotyping arrays, as differences in genotype quality can induce spurious findings in this setting[100].

## Box 2 | An analysis to identify CP effect loci for psychiatric disorders




To illustrate a realistic application of a meta-analysis of cross-phenotype (CP) effects, we provide examples from a study of psychiatric disorders. The Psychiatric Genomics Consortium (PGC) conducted a largescale meta-analysis in 61,220 cases and controls across five psychiatric disorders[7]: autism spectrum disorder, attention-deficit hyperactivity disorder, bipolar disorder, major depressive disorder and schizophrenia (see part **a** of the figure). As data were collected across dozens of studies in over 19 countries and were genotyped on different arrays, all individual-level raw data were subjected to the same quality-control measures and followed the same protocol for imputation to obtain a common set of single-nucleotide polymorphisms (SNPs). This step is essential for reducing biases in the data that can lead to spurious CP associations (BOX 1). In

addition, controls appearing in more than one study were randomly assigned to a control group for one of the phenotypes.

Univariate genome-wide association studies (GWASs) were carried out for each phenotype after combining individual-level data for each disorder. A fixed-effects meta-analysis was carried out on the summary statistics from the univariate GWAS to test for CP effects genome-wide and identified four genome-wide-significant SNPs: two are shown in forest plots in part **b** of the figure. In each forest plot, the effect size and 95% confidence interval are plotted for each individual phenotype and for the overall meta-analysed results ('all' in the figure). Fixed-effects meta-analysis was chosen because the power can be higher than random-effects analysis for situations in which effects are not substantially different.

A significant CP result indicates that the SNP is associated with at least one of the phenotypes but requires an additional step to identify which phenotypes are driving the association (note that most meta-analytical approaches require this step (TABLE 2)). To identify which of the five phenotypes were associated, the authors used a multinomial logistic regression model developed by Lee *et al*.[45] that allows comparisons between multiple CP-specific disease models and uses a model selection step to identify the best-fitting configuration of disorder-specific CP effects. The approach jointly models multiple categorical phenotypes and requires the availability of individual-level genotype data. The model selection technique found that the best-fit model indicated an effect on all five phenotypes for rs11191454 and an effect limited to schizophrenia and bipolar disorder for rs1024582.

In addition to identifying SNP-level CP effects, polygenic scoring analyses were conducted to assess the overall evidence for pleiotropy among these disorders using thousands of SNPs in aggregate. The results indicated significant genetic overlap among schizophrenia, bipolar disorder and major depressive disorder and also between autism spectrum disorders and schizophrenia, although to a lesser extent.

## Box 3 | Functional characterization of CP effects

For a functional understanding of how a cross-phenotype (CP)-effect locus contributes to disease, various computational and experimental steps are carried out. The starting point is typically a genetic variant that was initially identified by genome-wide association studies (GWASs) and has subsequently satisfied CP-effect criteria across multiple traits. Therefore, many characterization steps are common between standard GWAS variants and CP-effect variants. These variants are typically not causal because they are usually only tags of the true causal variant or variants.

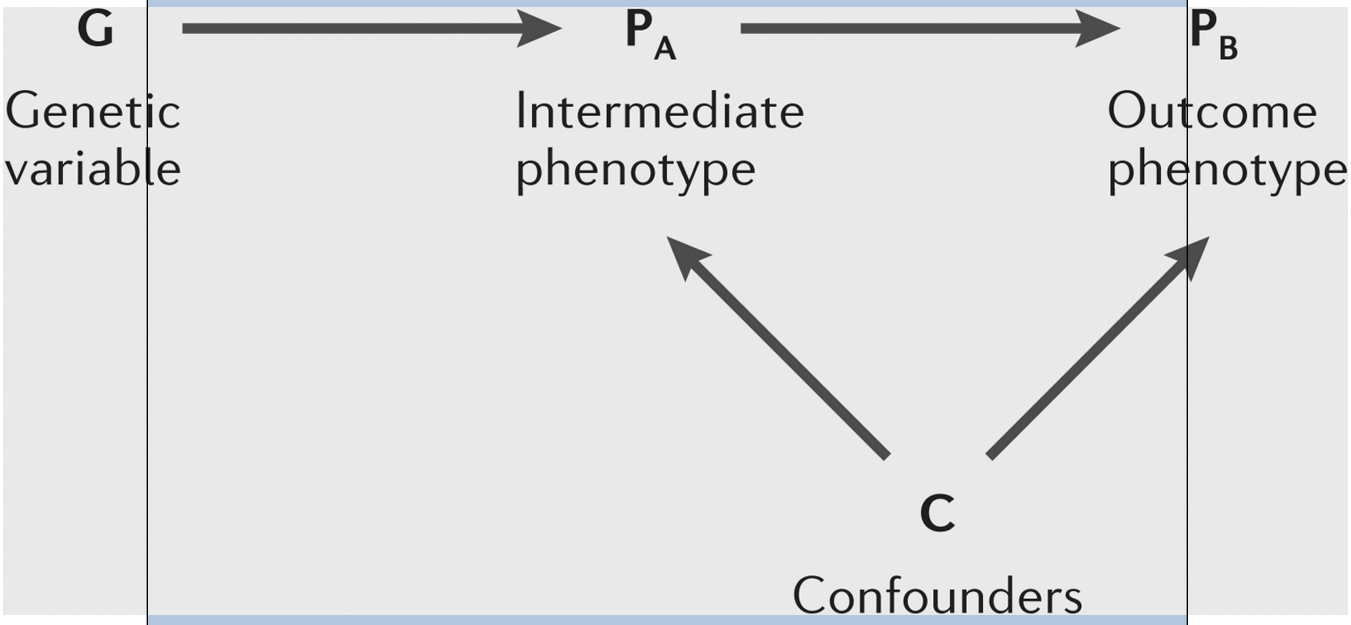
Sequencing and fine mapping are necessary steps for identifying causal variants in the region that affect more than one trait. After the causal variants have been identified, investigators are faced with the challenge of functionally characterizing the variants. Functional categories of the causal variants (such as missense or nonsense mutations in protein-coding genes) can provide crucial clues for characterizing the CP effects of genes. Various bioinformatics tools and databases[101] are available for predicting the deleterious, potentially disease-causing biomolecular effects of mutations on the basis of the functional category (such as PolyPhen[102] or SIFT[103]). Although most of these tools focus on the functional effects of either protein-coding or splice-site variants, mutations in non-protein-coding genes (such as microRNAs) or intergenic regulatory elements (such as enhancers) can result in the dysregulation of hundreds of target proteins and thus could have a major role (refer to the Encyclopedia of DNA Elements (ENCODE) project). It is also noteworthy that regulatory variants may confer tissue-specific effects on multiple genes[104], some of which could occur on different chromosomes (*trans*-effects[105]). Examination of expression quantitative trait locus (eQTL) data in a relevant tissue type can help to identify the tissue-specific regulatory changes caused by mutations[106,107], as demonstrated in the Genotype-Tissue Expression (GTEx) eQTL Project[108]. Thus, single variants can have distinct effects on different tissues.

The CP effect of a single variant can also occur when the gene is involved in multiple pathways or when it is involved in the same pathway but has a different phenotypic effect on the associated disorders. Public resources of canonical pathways, biological functions or protein–protein interaction data can be used to compare and contrast diverse biological roles of gene products as well as potential pathogenetic mechanisms underlying distinct disorders (for example, Pathguide)[109]. It is often informative to use a statistical strategy, such as multivariate pathway analysis, for identifying statistically enriched sets of biologically related genes for single disorders and comparing the implicated pathways in the functional characterization of the CP effect[110].

Finally, it is essential to validate the predicted CP effects of genetic variation on cellular physiology using experimental methods[104]. Typically, the molecular effects of a variant can be demonstrated using cultured cells *in vitro*, a knock-in or knockout strategy in animal models[111–113] or, more recently, *in vivo* genome editing[114,115]. Changes in protein expression, localization and mRNA transcription indicate the functional effects of the mutation. However, it should be noted that unless the replacement with the variant results in phenotypic changes that are directly related to the disorder, experimental validation of the functional effects does not necessarily imply causality to the disease in

humans[116]. Therefore, clarification of the CP effects requires that functional studies of gene mutation be carried out separately in pathogenic cell types that are relevant to the implicated disorders[117]. A successful example is a series of functional studies that verified the CP effects of an endogenous β-galactoside-binding protein galectin 3; the knockout mouse model of galectin 3 revealed that the deficiency of the protein leads to a concanavalin-A-induced hepatitis in the liver[118,119], whereas inhibition of galectin 3 expression suppresses tumour growth in human breast carcinoma cells[120–123].

## Box 4 | Mendelian randomization




Mendelian randomization is a form of instrumental variable analysis — a common approach in causal inference — that uses a genetic variable (G, the instrumental variable) to test whether an intermediate phenotype ($P_A$) causes another phenotype ($P_B$; see the figure)[73,74]. For example, Mendelian randomization was used to test whether the relationship between high-density lipoprotein cholesterol and myocardial infarction is causal[37]. To conduct a valid Mendelian randomization experiment, the following assumptions must be met[73,74]:

- Assumption 1: G (which is a single-nucleotide polymorphism (SNP) or a combination of multiple SNPs[124]) is robustly associated with $P_A$.
- Assumption 2: G is unrelated to C, which are confounding factors that bias the relationship between $P_A$ and $P_B$. In other words, there are no common causes of G and $P_B$.
- Assumption 3: G is related to $P_B$ only through its association with $P_A$.

If these assumptions are met, it is possible to test the hypothesis that $P_A$ causes $P_B$ and to derive an estimate of this relationship ($\beta_{IV}$) by using the regression coefficients for testing the association of G and $P_B$, and G and $P_A$[73]: $\beta_{IV} = \beta_{G,PB}/\beta_{G,PA}$.

The assumptions of Mendelian randomization are quite strong and thus the instrumental variable (G) must be carefully selected. Even small violations of the assumptions can result in severe bias[74]. Assumption 1 can easily be verified by selecting a G that is robustly associated with $P_A$. Assumption 2 generally holds because G is randomized at birth and thus is independent of non-genetic confounders and is not modified by the course of disease. However, population stratification could violate this assumption if ancestry is related both to G and to $P_B$. Assumption 3 explicitly assumes that G is not associated with $P_A$ and $P_B$ through biological pleiotropy, meaning that G is associated

with $P_B$ only through $P_A$ (that is, the association instead involves mediated pleiotropy) and that G is not associated with any unmeasured phenotype that is related to $P_B$. Knowledge about the causal nature of the association between G and $P_A$ can help to verify this assumption[74]. Additionally, using multiple different instrumental variables for different genes and showing consistent results can also help to rule out violations[73,74]. Although assumptions 2 and 3 cannot be empirically proved, there are several additional tests that can be used to try to falsify assumptions 2 and 3 and thus to minimize the chance of bias[74].

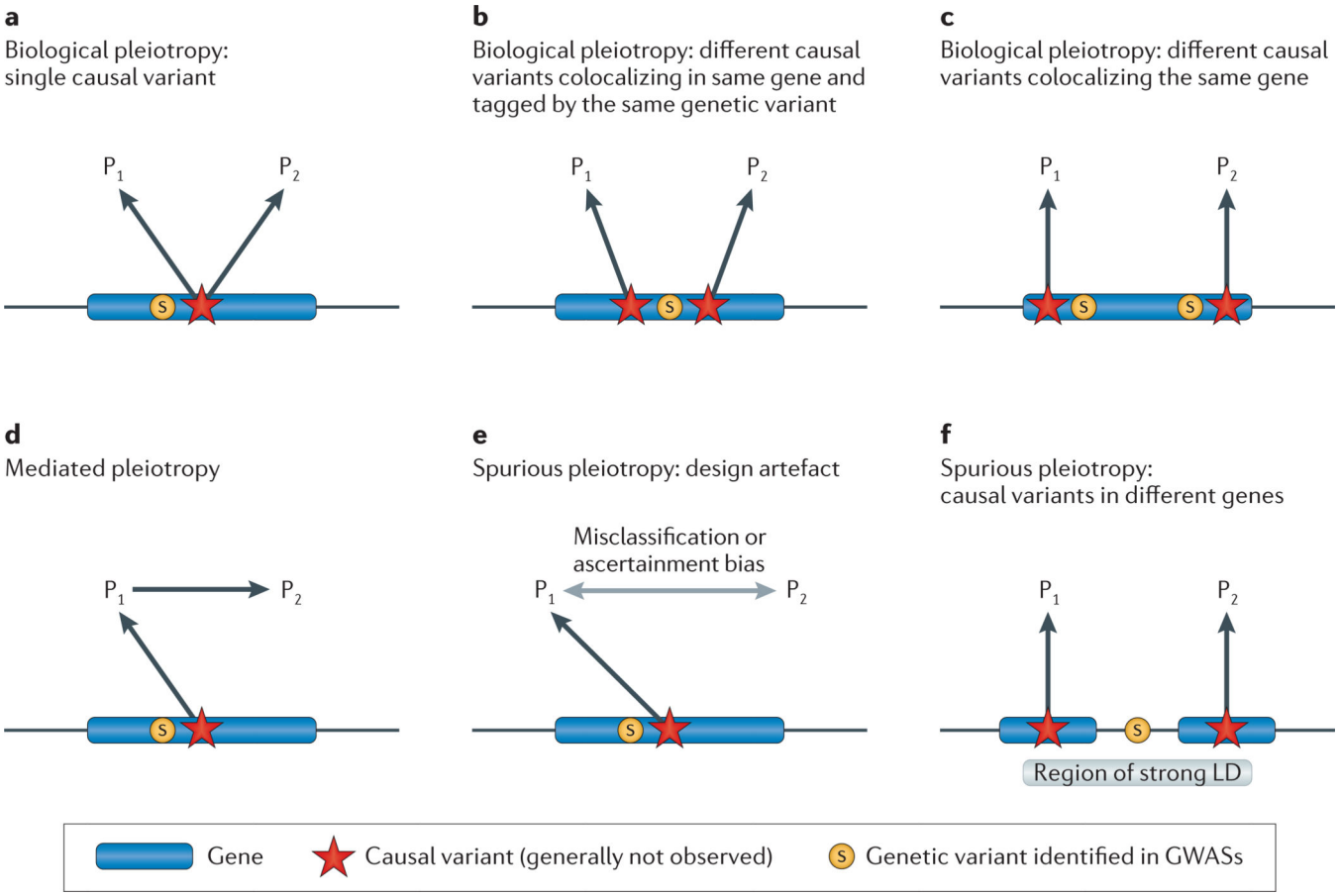



**Figure 1. Types of pleiotropy**

In each scenario, the observed genetic variant (S) is associated with phenotypes 1 and 2 ($P_1$ and $P_2$). We assume that the observed genetic variant is in linkage disequilibrium (LD) with a causal variant (red star) that affects one or more phenotypes. In some cases, the causal variant may be identified directly and the figures can be simplified accordingly. The various figures correspond to the unobserved underlying pleiotropic structure. **a** | Biological pleiotropy at the allelic level: the causal variant affects both phenotypes. **b** | Colocalizing association (biological pleiotropy): the observed genetic variant is in strong LD with two causal variants in the same gene that affect different phenotypes. **c** | Biological pleiotropy at the genic level: two independent causal variants in the same gene affect different phenotypes. **d** | Mediated pleiotropy: the causal variant affects $P_1$, which lies on the causal path to $P_2$, and thus an association occurs between the observed variant and both phenotypes. **e** | Spurious pleiotropy: the causal variant affects only $P_1$, but $P_2$ is enriched for $P_1$ owing to misclassification or ascertainment bias, and a spurious association occurs between the observed variant and the phenotype 2. **f** | Spurious pleiotropy: the observed variant is in LD with two causal variants in different genes that affect different phenotypes. GWAS, genome-wide association study.

## Table 1

CP associations in the literature

| Type | Locus | Phenotypes | Result | Refs |
|---|---|---|---|---|
| SNP (same direction of risk) | rs11209026 (*IL23R*) | Crohn's disease, ankylosing spondylitis, ulcerative colitis, psoriasis | The minor allele (A) of rs11209026 is protective for Crohn's disease, ankylosing spondylitis, ulcerative colitis and psoriasis | 125–128 |
| SNP (same direction of risk) | rs6983267 (8q24) | Prostate and colorectal cancer | The G allele increases risk for prostate cancer and colorectal cancer | 23,24 |
| SNP (different direction of risk) | rs12720356 (*TYK2*) | Crohn's disease and psoriasis | The G allele increases risk for Crohn's disease and decreases risk for psoriasis | 128,129 |
| Gene (different SNPs) | *DNAH11* | LDL cholesterol and multiple myeloma | rs12670798 is associated with LDL cholesterol and rs4487645 is associated with multiple myeloma | 130,131 |
| Gene (different SNPs) | *FTO* | BMI and melanoma | rs8050136 is associated with body mass index and rs16953002 is associated with melanoma | 17,18 |
| Region (different SNPs) | 9q21.3 | Coronary artery disease, glioma, intracranial aneurysm | rs4977574 is associatied with coronary artery disease, rs4977756 with glioma, rs1333040 with intracranial aneurysm | 19–22 |
| Copy number variation | 16p2.11 duplication | Schizophrenia, autism, intellectual disability, developmental delay, congenital malformations | CNV duplication increases risk for all five disorders | 26 |
| Copy number variation | 7q11.23 | Autism and Williams–Beuren syndrome | CNV deletion causes Williams–Beuren syndrome and *de novo* CNV duplication increases risk for autism | 132,133 |
| Pathway | Immune cell signalling | Autoimmune thyroid disease, coeliac disease, Crohn's disease, rheumatoid arthritis, systemic lupus erythematosus, T1D | Genes in this pathway have been implicated across six diseases | 34 |
| Polygenic scores | – | Schizophrenia and bipolar disorder | Schizophrenia and bipolar disorder share genetic factors that increase risk to both disorders | 28 |
| Genetic correlation | – | T2D and hypertension | Positive genetic correlation between T2D and hypertension suggests that shared genetic factors increase risk for both traits | 41 |

BMI, body mass index; CNV, copy number variant; CP, cross-phenotype; *DNAH11*, dynein, axonemal, heavy chain 11; *FTO*, fat mass and obesity associated; *IL23R*, interleukin 23 receptor; LDL, low-density lipoprotein; SNP, single-nucleotide polymorphism; T1D, type 1 diabetes; T2D, type 2 diabetes; *TYK2*, tyrosine kinase 2. This table provides some examples of different types of observed CP effects. These are illustrative examples and are not exhaustive; many additional CP associations have been published.

**Table 2**

Univariate approaches for detecting CP associations

| | Input | Explicit test of CP association | Allows effect heterogeneity | Types of phenotype (such as continuous or categorical) | Accommodates overlapping subjects | Combine data across multiple studies | Identify subset of associated phenotypes | Genetic variant versus region | Refs |
|---|---|---|---|---|---|---|---|---|---|
| **Fisher** | *P* value | No | Yes | Any | No | Yes | No | Variant | 56 |
| **CPMA** | *P* value | Yes | Yes | Any | No | Yes | No | Variant | 14 |
| **Fixed effects meta-analysis** | Effect estimate | No | No | Same type; need to standardize continuous phenotypes | No | Yes | No | Variant | 54,57,58// |
| **Random effects meta-analysis** | Effect estimate | No | Moderate level; not opposite effects | Same type; need to standardize continuous phenotypes | No | Yes | No | Variant | 54,57,58// |
| **Subset-based meta-analysis** | Effect estimate | No | Yes | Same type; need to standardize continuous phenotypes | No; offer extension to account for some overlap | Yes | Yes | Variant | 59 |
| **Extensions to O'Brien** | Effect estimate | No | Yes | Any | Yes; all subjects overlap* | No§ | No | Variant | 61,62 |
| **TATES** | *P* value | No | Yes | Any | Yes; all subjects overlap‡ | No§ | No | Variant | 63 |
| **PRIMe** | *P* value | No | Yes | Any | Yes | Yes | No | Region | 64 |

CP, cross-phenotype; CPMA, cross-phenotype meta-analysis; PRIMe, Pleiotropy Regional Identification Method; TATES, Trait-based Association Test that uses Extended Simes.

*Can accommodate values missing completely at random.

‡Can accommodate values missing completely at random and blockwise missingness.

§Can combine across multiple studies if all subjects have non-missing values for all phenotypes; TATES can accommodate situations in which a subset of studies have missing values for a subset of the phenotypes.

//References are given for meta-analytical methods typically used in genome-wide association studies.