**Abstract**

As our understanding of genetics has improved, genome-wide association studies (GWAS) have identified numerous variants associated with lifestyle behaviours and health outcomes. However, what is sometimes overlooked is the possibility that genetic variants identified in GWAS of disease might reflect the effect of modifiable risk factors as well as direct genetic effects. We discuss this possibility with illustrative examples from tobacco and alcohol research, in which genetic variants that predict behavioural phenotypes have been seen in GWAS of diseases known to be causally related to these behaviours. This consideration has implications for the interpretation of GWAS findings.

**Introduction**

The rapid growth in genome-wide association studies (GWAS) has resulted in the identification of common genetic variants associated with behavioural traits, from biomarker phenotypes that capture the downstream consequences of behaviour (e.g., body mass index [BMI]) [1] to the behaviours themselves (e.g., tobacco and alcohol use) [2]. While the success of GWAS has generated insights into the biological mechanisms underpinning these traits (see Box 1), it is less appreciated that it has also begun to tell us about the causal effects of modifiable or environmental influences on these traits. For example, a genetic variant at a locus containing the NPC1L1 gene is strongly associated with low-density lipoprotein (LDL) cholesterol level as well as with the risk of cardiovascular events. This is not because NPC1L1 is independently associated with cardiovascular problems, but simply because high cholesterol is a causal risk factor for the disease [3–5]. In other words, there are a number of cases where GWAS of disease outcomes have identified loci that capture modifiable risk factors rather than direct biological pathways. Here, we explain how this insight can inform the interpretation of GWAS results.
Box 1. **CHRNA5-A3-B4 and Cigarette Smoking**

The most robust finding to emerge from GWAS of smoking phenotypes is the association between the nicotinic receptor gene cluster **CHRNA5-A3-B4** on chromosome 15 (at 15q25) and smoking quantity. This gene cluster encodes three nicotinic acetylcholine receptor subunit proteins: alpha-5, alpha-3, and beta-4. An association between the non-synonymous variant rs16969968 in **CHRNA5** and nicotine dependence was first reported in 2007 in a candidate gene study [39], with the minor allele found to confer increased risk. The following year, the same locus (tagged by rs1051730 in **CHRNA3**, in high linkage disequilibrium with rs16969968) was found to be associated with smoking quantity, this time in a GWAS conducted by Thorgeirsson and colleagues [11]. This study also highlighted an association between rs1051730 and two smoking-related diseases, lung cancer and peripheral arterial disease. These initial findings renewed interest in **CHRNA5** and **CHRNA3** and the role played by these genes in nicotine dependence, and led on to a series of preclinical follow-up studies focused on determining the mechanism underlying the observed associations with smoking behaviour and disease. Functional research has demonstrated that the minor allele at rs16969968 (i.e., the risk variant for heavier smoking) is associated with a decreased maximal response to a nicotine agonist relative to the major allele in vitro [40]. Subsequent research using alpha-5 knockout mice has further illustrated the role that **CHRNA5** plays in determining response to nicotine. Using a nicotine self-administration paradigm, Fowler and colleagues [41] observed that knockout mice responded far more vigorously than wild-type mice for nicotine infusions at high doses and, unlike wild-types, did not self-titrate nicotine delivery. Deficient alpha-5 signalling attenuated the aversive effects of nicotine that would normally serve to limit its intake.

The association between the **CHRNA5-A3-B4** locus and smoking quantity [2,12,42–44], alongside other smoking-related phenotypes and diseases, has been replicated in numerous studies. GWAS of lung cancer [14] and chronic obstructive pulmonary disease [15] have also identified this locus, and follow-up candidate gene studies have suggested a role in bladder cancer [45], emphysema [46], and upper aerodigestive tract cancer [47], diseases for which smoking has already been recognised as a causal factor [48]. Some of these findings were used to argue that there is an independent effect of this locus on the disease, given evidence of residual association between variant and disease following adjustment for self-reported smoking quantity. However, this is likely due to the imprecision of self-report measures of heaviness of smoking and misclassification of smoking status. Studies using biomarkers of tobacco exposure have illustrated that rs1051730/rs16969968 accounts for a far greater proportion of variance in nicotine metabolite levels relative to self-report measures of daily tobacco consumption [13,49,50]. When we use the per allele effect of rs1051730 on cotinine levels, for example, to estimate the association between genotype and lung cancer risk, this accords with published data, which supports the conclusion that the effect of this locus on lung cancer risk is mediated largely, if not wholly, via level of tobacco exposure [13].

**Genetics and Causal Inference**

The relationship between phenocopy and genocopy (see Box 2) lies at the heart of the Mendelian randomization approach, which seeks to leverage genetic information to identify causal
The principles of Mendelian randomization have been described in detail elsewhere [6–8]. In brief, genetic variants are used as proxies (i.e., instrumental variables) for modifiable exposures. If the assumptions of Mendelian randomization hold, these proxies should not be associated with the factors that confound observational associations and will not be subject to reverse causation. This has become an increasingly popular technique for establishing whether an observational association between an exposure and an outcome is likely to be causal. However, while Mendelian randomization makes use of information obtained (principally) via GWAS, our argument is that the same reasoning can directly inform our interpretation of GWAS results.

Implementing Mendelian randomization techniques can be challenging, principally because single genetic variants (and even polygenic risk scores) typically capture only a small proportion of variance in the exposure of interest. An ideal instrument would exactly mimic the exposure of interest without being associated with confounding variables, but this is, of course, impossible in practice. Genetic variants are, therefore, generally weak instruments, meaning that very large sample sizes are required to attain adequate statistical power. Historically, datasets were required with information on genotype, outcome, and exposure of interest in order to run such studies. Methodological developments now allow the application of Mendelian randomization across two different samples if no single sample is available that includes data on genotype, the exposure, and the outcome.

Conventional Mendelian randomization uses genetic markers known to be associated with a modifiable exposure of interest, for which there is also a known observational association between the modifiable exposure and an outcome of interest. In an ideal situation, the association of the genotype with the outcome can be tested across strata of individuals who are positive or negative for the putative mediating exposure (e.g., ever-smokers versus never-smokers). If there is a causal effect of the exposure (e.g., smoking heaviness) that is being captured by the genotype, then an association of the genotype with the outcome should only be seen in the exposed group and not the unexposed group (see Fig 1) [9]. This is a special case of gene × environment (G × E) interaction, where both G and E are known, although it will not always be possible to stratify on the exposure, and stratification (which can be considered a form of statistical adjustment) can introduce other potential biases in certain circumstances.

Box 2. Genocopy and Phenocopy

The term “phenocopy” is attributed to Goldschmidt [51] and describes the situation where an environmental effect results in the same effect as that produced by a genetic variant. It is generally used to describe diseases that are similar to some genetic syndrome but that can also be caused by environmental exposures. The term “genocopy,” attributed to Schmalhausen [52], is essentially the reverse of phenocopy and describes the situation in which a genetic variant produces an outcome that could equally be produced by an environmental exposure. The critical point is that genetic and environmental causes of disease can be seen as essentially equivalent; as Goldschmidt wrote in 1938, “different causes produce the same end effect, presumably by changing the same developmental processes in an identical way” [51]. More recently, Zuckerkandl and Villet have argued that “no doubt all environmental effects can be mimicked by one or several mutations,” again suggesting that genetic and environmental influences can be regarded as both equivalent and interchangeable [53].
Nevertheless, if differences in the magnitude of association observed in exposed and unexposed groups are very large, this is convincing evidence of a causal pathway via the exposure. Therefore, a logical extension of Mendelian randomization is that any GWAS will, in principle, identify exposures that are causally associated with the outcome of interest in that GWAS.

We illustrate the application of Mendelian randomization reasoning to GWAS data using the example of two behavioural phenotypes—tobacco and alcohol use. These are phenotypes for which GWAS has identified a number of associated loci. A number of disease outcomes are also known to be associated with tobacco and alcohol use, and therefore serve as proof of principle for our argument that GWAS can identify the effects of modifiable exposures.

Cigarettes and Alcohol

The strongest genetic signal for tobacco-use phenotypes identified via GWAS is located in a gene cluster on chromosome 15 containing the CHRNA5, CHRNA3, and CHRNA4 genes (CHRNA5-A3-B4), which encode the alpha-5, alpha-3, and beta-4 nicotinic acetylcholine receptor subunits, respectively. Each additional copy of the minor risk allele at this locus is associated with one extra cigarette smoked per day. The locus [11] accounts for approximately 1% of the variation in cigarette consumption in daily smokers [12] and approximately 4% of the variation in cotinine levels, the primary metabolite of nicotine and a more precise biomarker of exposure [13]. The same locus has been identified in GWAS of lung cancer [11,14], peripheral arterial disease [11], and chronic obstructive pulmonary disease [15]. One parsimonious interpretation of these results is that these are all diseases for which cigarette smoking is
a strong, causal risk factor. Indeed, the effect of smoking on these outcomes is sufficiently strong that variants associated with heaviness of smoking achieve genome-wide significance even in unstratified GWAS (i.e., where smokers and never-smokers are not considered separately). When stratified, one should see the association only in ever-smokers and not in never-smokers (see Box 3) [16–18], although due to misclassification (i.e., misreporting of smoking status), this is not always the case.

A subtler example arises from the association of \textit{ALDH2} with alcohol consumption. This gene encodes aldehyde dehydrogenase, an enzyme responsible for metabolizing acetaldehyde (a metabolite of alcohol) to acetic acid. When less of this enzyme is present, acetaldehyde can

---

**Box 3. Stratification in Mendelian Randomization**

For behavioural exposures such as tobacco or alcohol use, stratification on the exposure of interest can be a powerful means of testing the pleiotropy assumption that is central to Mendelian randomization. While endogenous exposures (e.g., cholesterol levels) can never be zero, only higher or lower in different individuals, behavioural exposures are generally limited to a subset of the population (e.g., smokers). In principle, a genetic variant associated with heaviness of smoking should be associated with an outcome of interest (e.g., BMI) only among those exposed to this putative causal agent (i.e., ever-smokers) and not those unexposed (i.e., never-smokers). Whether this association differs between strata can be assessed using an interaction test.

However, stratification on a common effect can introduce collider bias [10,54], which can result in a spurious correlation between otherwise independent variables (Fig 2A). In the case of the \textit{CHRNA5-A3-B4} variants used in Mendelian randomization analyses of smoking, the assumption is that these variants are principally associated with heaviness of smoking rather than smoking status, in which case the risk of collider bias is reduced, as smoking initiation is not a common effect (Fig 2B). However, if these variants are shown to be associated with smoking initiation [55], this risk would be increased. Stratification does not always introduce the risk collider bias—for example, in a Mendelian randomization analysis of alcohol consumption and blood pressure, the analysis was stratified by participant sex due to differences in alcohol consumption among men and women in East Asian populations [23]. This does not introduce the possibility of collider bias because sex cannot be an effect of the genetic variant in question; sex is determined by a different genetic variation, which is inherited independently of other variants, and sex cannot be an effect of blood pressure (Fig 2C).

It is also worth remembering that the environmental exposure that is used for stratification is subject to the usual problems of confounding. Keller has argued that many gene × environment interaction studies do not adequately control for potential confounders because they do not include covariate × gene and covariate × environment interaction terms [56]. For example, \textit{ADH1B} genotype shows clear association with risk of upper aerodigestive cancer among alcohol drinkers, but not non-drinkers, consistent with a causal effect of alcohol consumption (Fig 3) [14]. However, a similar (albeit weaker) pattern is observed when stratification is based on smoking status rather than alcohol consumption, because these exposures are correlated. In these cases, the interaction will be stronger for the causal factor (i.e., when stratification is based on drinking status rather than smoking status).
build up after alcohol consumption, leading to unpleasant side effects. Therefore, the minor allele is robustly associated with reduced alcohol consumption [19]. The frequency of the minor allele at the \textit{ALDH2} locus is very low in samples of European ancestry but is relatively common in samples of East Asian ancestry. It is, therefore, only associated with alcohol consumption in the latter population. As a result, in GWAS of high blood pressure, \textit{ALDH2} was not identified in studies that recruited predominantly European samples [20] but was identified in studies that recruited East Asian samples (once genotyping chips that adequately tagged the \textit{ALDH2} locus were used) [21,22]. This confirms the results of Mendelian randomization analyses of alcohol consumption and blood pressure conducted prior to these later GWAS [23]. In other words, alcohol consumption causes high blood pressure, and this is detected in GWAS of high blood pressure, but only when tested in populations in which the variants associated with alcohol consumption are sufficiently common. This provides strong evidence that the identification of this locus in the GWAS is due to a causal effect of alcohol consumption rather than being due to shared genetic aetiology or to pleiotropy. Similarly, \textit{ALDH2} has emerged in GWAS of esophageal cancer in East Asian samples [24], confirming earlier Mendelian randomization analyses [25].
Behavioural traits such as tobacco and alcohol use can be regarded as intermediate traits, which are under a degree of genetic influence but which are themselves direct causal agents influencing various health outcomes. However, a critical difference between these and more usual intermediate phenotypes (such as LDL cholesterol) is that whereas both may be direct causal agents and amenable to intervention for therapeutic benefit, the former may be entirely absent (i.e., non-smokers, non-drinkers), whereas the latter cannot be (i.e., no one has a cholesterol level of zero). Genetic variants may influence whether or not someone smokes or drinks, or how much they smoke or drink, or both.

Implications

As GWAS of disease outcomes are carried out on increasingly large samples, more loci will be identified, promising to deliver insights into underlying biological mechanisms. However, as we have seen, it will become increasingly important to also consider whether these associations reflect effects of modifiable exposures. This will require the triangulation of evidence across GWAS of disease outcomes and GWAS of behavioural phenotypes to determine the cases in which signals identified for behavioural phenotypes are the same as those identified for disease phenotypes. Unfortunately, this approach is hampered at present by the relative lack of GWAS of behavioural phenotypes—while we have identified a number of variants associated with tobacco and (to a lesser extent) alcohol use, as well as obesity, this is not yet the case for exposures such as cannabis use. Nevertheless, this situation is rapidly changing—for example, there are now several variants that have been shown to be associated with caffeine consumption [26]. It is also worth noting that both the CHRNA5-A3-B4 and ALDH2 loci were initially identified in candidate gene studies.
Already, intriguing hints are emerging that larger GWAS are beginning to identify potential environmental or behavioural causes of disease. A recent GWAS led by the Psychiatric Genomics Consortium identified 108 loci associated with schizophrenia [27]. One locus that reached genome-wide significance is located in the \textit{CHRNA5-A3-B4} gene cluster on chromosome 15, which, as we have seen, has been consistently shown to contain multiple loci strongly associated with heaviness of smoking [2] among cigarette smokers. There are two possible explanations for this finding. One is that there may be genetic variants in this region that independently influence both heaviness of smoking and schizophrenia risk (i.e., genetic pleiotropy). The other is that this signal captures a causal effect of cigarette smoking on schizophrenia (reflecting a dose–response relationship among the smokers in the study). Again, there is a precedent for this pattern of results: the same region was shown to be associated with lung cancer risk [11], but it is likely that this effect operates entirely via cigarette smoking [13].

Critically, while the identification of 108 loci associated with schizophrenia was rightly heralded as a breakthrough in our understanding of the genetic determinants of schizophrenia, very little was made of this potentially vital insight. If smoking is indeed a causal risk factor for schizophrenia, then this has immediate and dramatic implications for public health, prevention, and treatment. Intriguingly, since the publication of these results, several other studies have been published that also support a causal role for smoking in schizophrenia and psychosis [28–31]. It is notable that one study reports a stratified analysis that suggests an association of \textit{CHRNA5-A3-B4} genotype with antipsychotic medication prescription (as a proxy of psychotic illness) in ever-smokers but not in never-smokers [32]. This is analogous to the case of \textit{CHRNA5-A3-B4} genotype and lung cancer risk, although the evidence in relation to schizophrenia is currently only suggestive.

As GWAS of other behavioural phenotypes such as personality and intelligence emerge, it will be interesting to see whether variants known to influence tobacco or alcohol use emerge, given the strong observational associations between these phenotypes. At the same time, GWAS of other behavioural phenotypes such as cannabis use will in due course provide loci that may signal causal effects of these behaviours on a range of other outcomes (notably schizophrenia).

**Identifying Causal Pathways**

For any locus identified via GWAS, we need to consider whether this reflects a potential modifiable risk factor. However, it is difficult to exclude the possibility that this locus is independently associated with both a modifiable risk factor and the disease outcome directly. For example, a recent study found an association between a polygenic risk score for schizophrenia (combining multiple variants identified with genome-wide significance into a single risk score) and cannabis use [32]. The authors concluded that this indicates that some of the association between schizophrenia and cannabis is due to a shared genetic aetiology. However, an alternative explanation could be that genetic predisposition to schizophrenia (and behaviours associated with this) increases the risk of cannabis use. Here, the distinction between mediated and biological pleiotropy is useful—the former refers to the genetic influence on the outcome operating via an exposure or intermediate phenotype, while the latter refers to a direct and independent genetic influence on both the exposure and the outcome [33]. Mediated pleiotropy is a single process leading to a cascade of downstream events, ultimately leading to a distal outcome. In this way, genetic variation at the \textit{FTO} locus influences BMI and, in turn, blood pressure, hypertension, coronary heart disease, and so on [33]. While statistical adjustment (e.g., for BMI) can help dissect these pathways, this can be problematic where residual associations may exist due to measurement error, such as in the case of \textit{CHRNA5-A3-B4}, smoking, and...
lung cancer risk [13]. Biological pleiotropy is more problematic and renders causal inference difficult.

A hierarchy of approaches supports stronger causal inference regarding the role of modifiable exposures on disease outcomes (see Table 1). Ultimately, what is required is a triangulation of evidence using these different approaches, ranging from whole genome methods to more focused analyses, to determine whether the results obtained using these different methods align consistently [34]. First, genetic correlation [35,36] can be used to identify shared genetic influences (e.g., cannabis use and schizophrenia). This approach allows all genotyped common variants to be interrogated, with correlations with modifiable exposures suggestive of possible causality. Second, conventional Mendelian randomization analyses (using single variants or polygenic risk scores) can be used to establish evidence that genetic proxies for a modifiable exposure of interest (e.g., cannabis use) associate with the outcome thought to be influenced by the exposure (e.g., schizophrenia) [33]. Single variant approaches are appropriate when the genetic variants play a known and relatively specific role in the pathway of interest (e.g., \textit{ALDH2} and alcohol consumption), but these will capture a smaller proportion of the variance in the exposure than polygenic risk scores. Third, when adequate genetic variants have been identified for both the exposure and the outcome, bidirectional Mendelian randomization can be used to determine with greater confidence the likely direction of any causal relationship.

<table>
<thead>
<tr>
<th>Strength of evidence (low &gt; high)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic correlation</td>
<td>This method estimates genetic correlation using GWAS summary statistics, using properties of linkage disequilibrium to allow for rapid screening for correlations among a diverse set of traits without the need for individual level data. However, this approach is still subject to genetic confounding (pleiotropy) and misclassification bias and requires larger samples than methods that use individual data. A well-powered null finding would argue against a causal association between exposure and outcome. However, direction of causation cannot be identified.</td>
</tr>
<tr>
<td>Polygenic risk score association</td>
<td>Polygenic risk scores can be derived where there are multiple variants identified with genome-wide significance for a trait or disease. These can be weighted to represent the proportion of the variance in the risk factor that they explain, and used as a proxy for an exposure to investigate associations of interest. The use of a risk score allows for a larger proportion of the variance to be explained, although it is very likely it will increase the risk of pleiotropy.</td>
</tr>
<tr>
<td>Bidirectional Mendelian randomization with polygenic risk scores</td>
<td>If polygenic risk scores are available for both the exposure and outcome of interest, associations can be investigated in both directions, which may provide evidence in support of an association in a particular causal direction.</td>
</tr>
<tr>
<td>Mendelian randomization sensitivity analyses</td>
<td>Mendelian randomization Egger regression extends the basic Mendelian randomization method by meta-analysing the SNP outcome association from each individual SNP that is associated with the exposure. This treats each SNP as akin to a small study in a traditional meta-analysis. Regression analysis, allowing variation in the intercept, means it is able to provide an estimate of the extent to which genetic pleiotropy has an impact on the causal estimates from Mendelian randomization analyses. Kang median instrument analysis has been shown to identify causal effects as long as fewer than 50% of instruments are invalid, without requiring knowledge of which instruments are invalid. It also allows identification of when this 50% threshold is violated.</td>
</tr>
</tbody>
</table>

*PLOS Genetics | DOI:10.1371/journal.pgen.1005765 February 11, 2016*
Fourth, a range of sensitivity analyses exist that can inform the interpretation of the findings, such as the extent to which (biological) pleiotropy has an impact on the causal estimates derived from conventional Mendelian randomization methods. This may be particularly relevant when polygenic risk scores comprising variants that act on a range of biological pathways are used. These methods include Mendelian randomization Egger regression [37] and the Kang median instrument approach [38]. Those relationships that survive this hierarchy of approaches are strong candidates for further interrogation in mechanistic or experimental studies.

The approaches described here can also be informative with respect to null results. If a modifiable exposure is under genetic influence and is also causally related to a disease outcome, we would expect to eventually see genetic variants associated with the exposure emerge in a GWAS of the outcome, given sufficient sample size. If this is not seen, this suggests that there may be no causal pathway operating (or that any causal relationship is very weak). Of course, interpreting null results must be done cautiously, particularly in cases where the prevalence of the modifiable exposure or the minor allele frequency differs across populations. Current GWAS cannot control for these sources of heterogeneity, which may impact the power of GWAS to identify modifiable exposures in the way we have described. Cross-contextual comparisons (e.g., across GWAS conducted in different populations) may be informative in these cases.

Conclusion

As we run larger and larger GWAS, some of the signals that emerge may turn out to reflect the action of modifiable (e.g., environmental or behavioural) exposures, rather than more direct biological effects. At present, what is likely to be required to understand these pathways is a two-step approach in which initial GWAS findings are interrogated further in studies in which detailed phenotype information is available. At present, this is not always possible—for example, a lack of smoking status information in the studies contributing to the recent schizophrenia GWAS means it is not possible to test the possible causal effect of smoking in a stratified analysis. However, as large, richly phenotyped cohort studies (e.g., UK Biobank) emerge, it will become possible to identify modifiable exposures from genetic data and to dissect those pathways within the same cohort. Here, “modifiable” can refer to substance use, but also to factors such as cholesterol or metabolite levels or blood pressure, which are directly influenced by lifestyle choices. A failure to appreciate this point will hamper our ability to translate the results of GWAS into health benefits, by focusing attention on possible biological pathways when, in fact, the target for intervention could be a modifiable environmental or behavioural exposure. We also need to be cautious when using statistical adjustment to test whether a genetic variant operates entirely via the supposed intermediate behavioural pathway. Sometimes, the most parsimonious explanation (e.g., smoking causes lung cancer) is the best one.

References

4. Ference BA, Majeed F, Penumetcha R, Flack JM, Brook RD. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by
polymorphisms in NPC1L1, HMGCR, or both: a 2 x 2 factorial mendelian randomization study. Journal of the American College of Cardiology. 2015; 65:1552–1561. doi:10.1016/j.jacc.2015.02.020 PMID: 25770315


