nature genetics

Detection and interpretation of shared genetic influences on 42 human traits

Joseph K Pickrell[1,2], Tomaz Berisa[1], Jimmy Z Liu[1], Laure Ségurel[3], Joyce Y Tung[4] & David A Hinds[4]

We performed a scan for genetic variants associated with multiple phenotypes by comparing large genome-wide association studies (GWAS) of 42 traits or diseases. We identified 341 loci (at a false discovery rate of 10%) associated with multiple traits. Several loci are associated with multiple phenotypes; for example, a nonsynonymous variant in the zinc transporter *SLC39A8* influences seven of the traits, including risk of schizophrenia (rs13107325: log-transformed odds ratio (log OR) = 0.15, $P = 2 \times 10^{-12}$) and Parkinson disease (log OR = −0.15, $P = 1.6 \times 10^{-7}$), among others. Second, we used these loci to identify traits that have multiple genetic causes in common. For example, variants associated with increased risk of schizophrenia also tended to be associated with increased risk of inflammatory bowel disease. Finally, we developed a method to identify pairs of traits that show evidence of a causal relationship. For example, we show evidence that increased body mass index causally increases triglyceride levels.

The observation that a genetic variant affects multiple phenotypes (a phenomenon often called 'pleiotropy' (refs. 1–3), although we will not use this term) is informative in a number of applications. One such application is learning about the molecular function of a gene. For example, men with cystic fibrosis (primarily known as a lung disease) are often infertile because of congenital absence of the vas deferens; this is evidence of a shared role for the CFTR protein in lung function and the development of reproductive organs[4]. Another application is learning about the causal relationships between traits. For example, individuals with congenital hypercholesterolemia also have elevated risk of heart disease[5]; this is now interpreted as evidence that changes in lipid levels causally influence heart disease risk[6].

In these two applications, the same observation—that a genetic variant influences two traits—is interpreted in fundamentally different ways depending on known aspects of biology. In the first case, a genetic variant influences two phenotypes through independent physiological mechanisms (graphically, $P_1 \leftarrow G \rightarrow P_2$, if G represents the genotype, P_1 the first phenotype, and P_2 the second phenotype and the arrows represent causal relationships[7]), whereas, in the second case, the effect of the variant on the second trait is mediated through its effect on the first trait, $G \rightarrow P_1 \rightarrow P_2$. In some situations, knowing which interpretation of the observation to prefer is simple: for example, it seems difficult to imagine how the reproductive and lung phenotypes of a *CFTR* mutation could be related in a causal chain. In other situations, interpretation is considerably more challenging. For example, the causal connections between various lipid phenotypes and heart disease have been debated for decades (for example, see ref. 8).

As the number of reliable associations between genetic variants and various phenotypes has grown over the last decade[9], these issues have received increasing attention. A number of recent studies have identified genetic variants associated with multiple traits[10–20]; in general, these associations are interpreted as most plausibly due to the independent effects of a genetic variant on different aspects of physiology. For example, a genetic variant in *LGR4* is associated with bone mineral density (BMD), age at menarche, and risk of gallbladder cancer[16], presumably owing to effects mediated through different tissues.

There has also been increasing interest in the alternative, causal framework for interpreting genetic variants that influence multiple phenotypes, which has been formalized under the name 'Mendelian randomization' (refs. 21–23). Mendelian randomization has been used to provide evidence for (or against) a causal role for various clinical variables in disease etiology[24–30]. For example, genetic variants associated with body mass index (BMI) are also associated with type 2 diabetes[27]; this is consistent with a causal role for weight gain in the etiology of diabetes.

Thus far, most studies of multiple traits have been performed across the genome on groups of traits already known or hypothesized to be related[10,31–33] or via testing small sets of variants for effects on a wide range of traits[20,34]. We aimed to systematically perform a genome-wide search for genetic variants that influence pairs of traits and then to interpret these associations in light of the causal and non-causal models described above. In this paper, we describe the results of such a search using large GWAS of 42 traits.

RESULTS

We assembled summary statistics from 43 GWAS of 42 traits or diseases performed in individuals of European descent (**Table 1**; 2 of these GWAS were for age at menarche). These studies span a wide range of phenotypes, from anthropometric traits (for example, height, BMI, and nose size) to neurological disease (for example, Alzheimer disease and Parkinson disease) to susceptibility to infection (for example, childhood ear infections and tonsillectomy). Seventeen

[1]New York Genome Center, New York, New York, USA. [2]Department of Biological Sciences, Columbia University, New York, New York, USA. [3]UMR 7206 Eco-Anthropologie et Ethnobiologie, CNRS, MNHN, Université Paris Diderot, Sorbonne Paris Cité, Paris, France. [4]23andMe, Inc., Mountain View, California, USA. Correspondence should be addressed to J.K.P. (jkpickrell@nygenome.org).

Received 15 June 2015; accepted 20 April 2016; published online 16 May 2016; doi:10.1038/ng.3570

of these GWAS were performed by the personal genomics company 23andMe and have not previously been reported (for details of these studies, see **Supplementary Data 1–17**). For studies that were not done using imputation to all variants in Phase 1 of the 1000 Genomes Project[35], we performed imputation at the level of summary statistics with ImpG v1.0 (ref. 36). We estimated the approximate number of independent associated variants (at a false discovery rate (FDR) of 10%) in each study using fgwas v.0.3.6 (ref. 37). The number of associations ranged from around 5 (for age at voice drop in men) to over 500 (for height).

Identification of genetic variants that influence pairs of traits

We first aimed to identify genetic variants that influence pairs of traits. To do this, we developed a statistical model (extending that

Table 1 Phenotypes analyzed in this study

Phenotype	Abbreviation	Data source	Approx. number of loci	Approx. number of participants, in thousands (cases/controls, if applicable)
Neurological phenotypes				
Alzheimer disease	AD	Ref. 75	11	17/37
Migraine	MIGR	23andMe	37	53/231
Parkinson disease	PD	23andMe	43	10/325
Photic sneeze reflex	PS	23andMe	66	32/67
Schizophrenia	SCZ	Ref. 59	222	34/46
Anthropometric and social traits				
Beighton hypermobility	BHM	23andMe	18	64
Breast size	CUP	23andMe	14	34
Body mass index	BMI	Ref. 72	30	240
Bone mineral density (femoral neck)	FNBMD	Ref. 17	19	33
Bone mineral density (lumbar spine)	LSBMD	Ref. 17	21	32
Chin dimples	DIMP	23andMe	57	58/13
Educational attainment	EDU	Ref. 76	93	294
Height	HEIGHT	Ref. 71	584	253
Male-pattern baldness	MPB	23andMe	49	9/8
Nearsightedness	NST	23andMe	183	106/86
Nose size	NOSE	23andMe	13	67
Waist–hip ratio	WHR	Ref. 77	13	143
Unibrow	UB	23andMe	61	69
Immune-related traits				
Any allergies	ALL	23andMe	43	67/114
Asthma	ATH	23andMe	35	28/129
Childhood ear infections	CEI	23andMe	15	47/75
Crohn's disease	CD	Ref. 78	61	6/15
Hypothyroidism	HTHY	23andMe	30	18/117
Rheumatoid arthritis	RA	Ref. 79	74	14/44
Tonsillectomy	TS	23andMe	48	60/113
Ulcerative colitis	UC	Ref. 78	42	7/21
Metabolic phenotypes				
Age at menarche	AAM	Ref. 43	70	133
Age at menarche (23andMe)	AAM (23)	23andMe	55	77
Age at voice drop	AVD	23andMe	5	56
Coronary artery disease	CAD	Ref. 45	11	22/65
Type 2 diabetes	T2D	Ref. 80	11	12/57
Fasting glucose	FG	Ref. 81	15	58
Low-density lipoproteins	LDL	Ref. 82	41	85
High-density lipoproteins	HDL	Ref. 82	46	89
Triglycerides	TG	Ref. 82	31	86
Total cholesterol	TC	Ref. 82	53	89
Hematopoietic traits				
Hemoglobin	HB	Ref. 83	16	51
Mean cell hemoglobin concentration	MCHC	Ref. 83	15	46
Mean red blood cell volume	MCV	Ref. 83	42	48
Packed red blood cell volume	PCV	Ref. 83	13	44
Red blood cell count	RBC	Ref. 83	25	45
Platelet count	PLT	Ref. 84	50	44
Mean platelet volume	MPV	Ref. 84	29	17

For each study, we show the name of the phenotype, the abbreviation that is used throughout this paper, the data source, the number of independent autosomal loci identified at an FDR of 10%, and the number of participants in the study. For studies where the data source is 23andMe, a complete description of the GWAS is presented in the **Supplementary Note** and **Supplementary Data 1–17**.

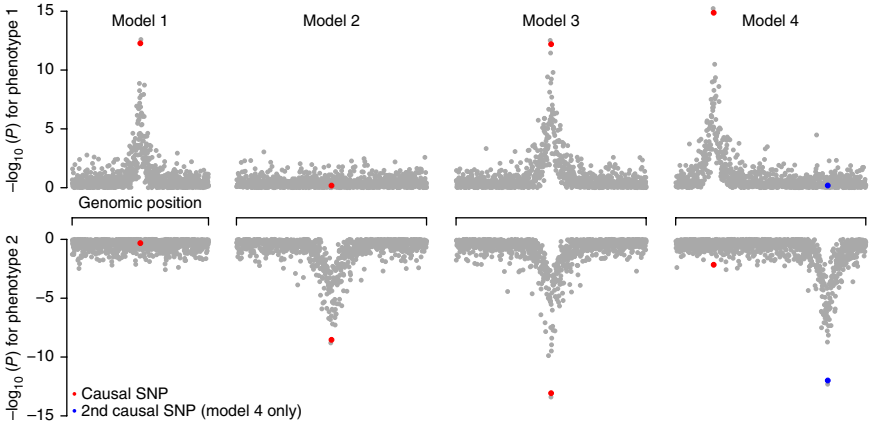

Figure 1 Schematic of the different models considered for a given genomic region and two GWAS. We divide the genome into approximately independent blocks (Online Methods) and estimate the proportion of blocks that fit into the shown patterns. The null model with no associations is not shown. Each point represents a single genetic variant.

used by Giambartolomei *et al.*[38]) to estimate the probability that a given genomic region (i) contains a genetic variant that influences the first trait (model 1); (ii) contains a genetic variant that influences the second trait (model 2); (iii) contains a genetic variant that influences both traits (model 3); or (iv) contains both a genetic variant that influences the first trait and a separate genetic variant that influences the second trait (model 4) (**Fig. 1**). The input to the model is the set of summary statistics (effect size estimates and standard errors) for each SNP in the genome on each of the two phenotypes, and (if the two GWAS were performed on overlapping sets of individuals) the expected correlation in the summary statistics due to correlation between the phenotypes. We can then fit the following log likelihood function

$$l(\Theta|D) = \sum_{i=1}^{M} \ln\left(\Pi_0 + \sum_{j=1}^{4} \pi_j \mathrm{RBF}_i^{(j)}\right)$$

where D is the data, M is the number of approximately independent blocks in the genome, Π_0 is the prior probability that a region contains no genetic variants that influence either trait, Π_1, Π_2, Π_3, and Π_4 represent the prior probabilities of the four models described above, Θ is the set of all five Π parameters, and $\mathrm{RBF}_i^{(j)}$ is the regional Bayes factor measuring the support for model j in genomic region i (see the **Supplementary Note** for details). In the presence of missing data, we consider only the subset of SNPs with data in both studies; if the causal SNP is not present, this acts to reduce power to detect a shared effect[38]. In fitting this model, we estimate the prior parameters and the posterior probability of each model for each region of the genome (for numerical stability, in practice, we penalize the estimates of the prior parameters and so obtain maximum a posteriori estimates). We were mainly interested in the estimated prior probability that each genomic region contains a variant that influences both traits ($\widehat{\Pi_3}$) and the corresponding posterior probabilities for each genomic region.

Several caveats of this method are worth mentioning. First, note that the estimate $\widehat{\Pi_3}$ is best thought of as the proportion of genomic regions that detectably influence both traits—if one study is small and underpowered, this estimate will necessarily be zero. This approach contrasts with methods that aim to provide unbiased estimates of the 'genetic correlation' between traits, which do not depend on sample size[39–41]. Second, in general, it is not possible to distinguish a single causal variant that influences both traits (model 3 in **Fig. 1**) from two separate causal variants (model 4 in **Fig. 1**) in the presence of strong linkage disequilibrium (LD) between the causal variants. For any individual genomic region discussed below, the possibility of two highly correlated causal variants must be considered as an alternative possibility in the absence of functional follow-up. (Indeed, this latter possibility appears to be common in quantitative trait locus studies performed in model organisms[42].)

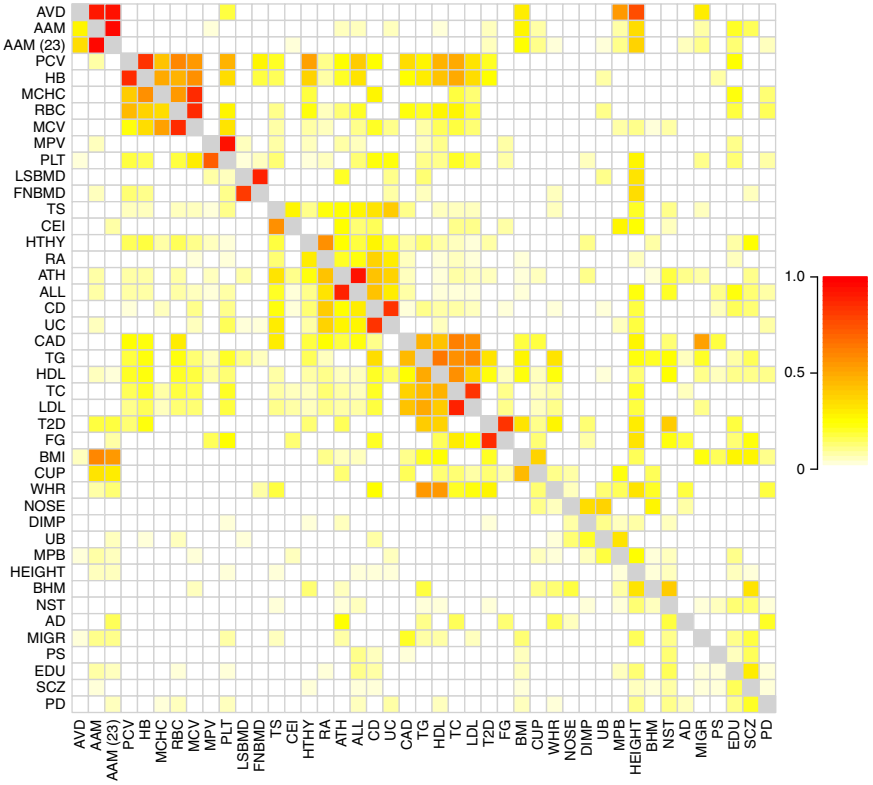

Figure 2 Heat map showing patterns of overlap between traits. Each square [*i*,*j*] shows the maximum a posteriori estimate of the proportion of genetic variants that influence trait *i* that also influence trait *j*, where *i* indexes rows and *j* indexes columns. Note that this is not symmetric. Darker colors represent larger proportions. Colors are shown for all pairs of traits that had at least one associated region in the set of 341 identified loci; all other pairs are set to white. Phenotypes were clustered by hierarchical clustering in R (ref. 74). Abbreviations are defined in **Table 1**.

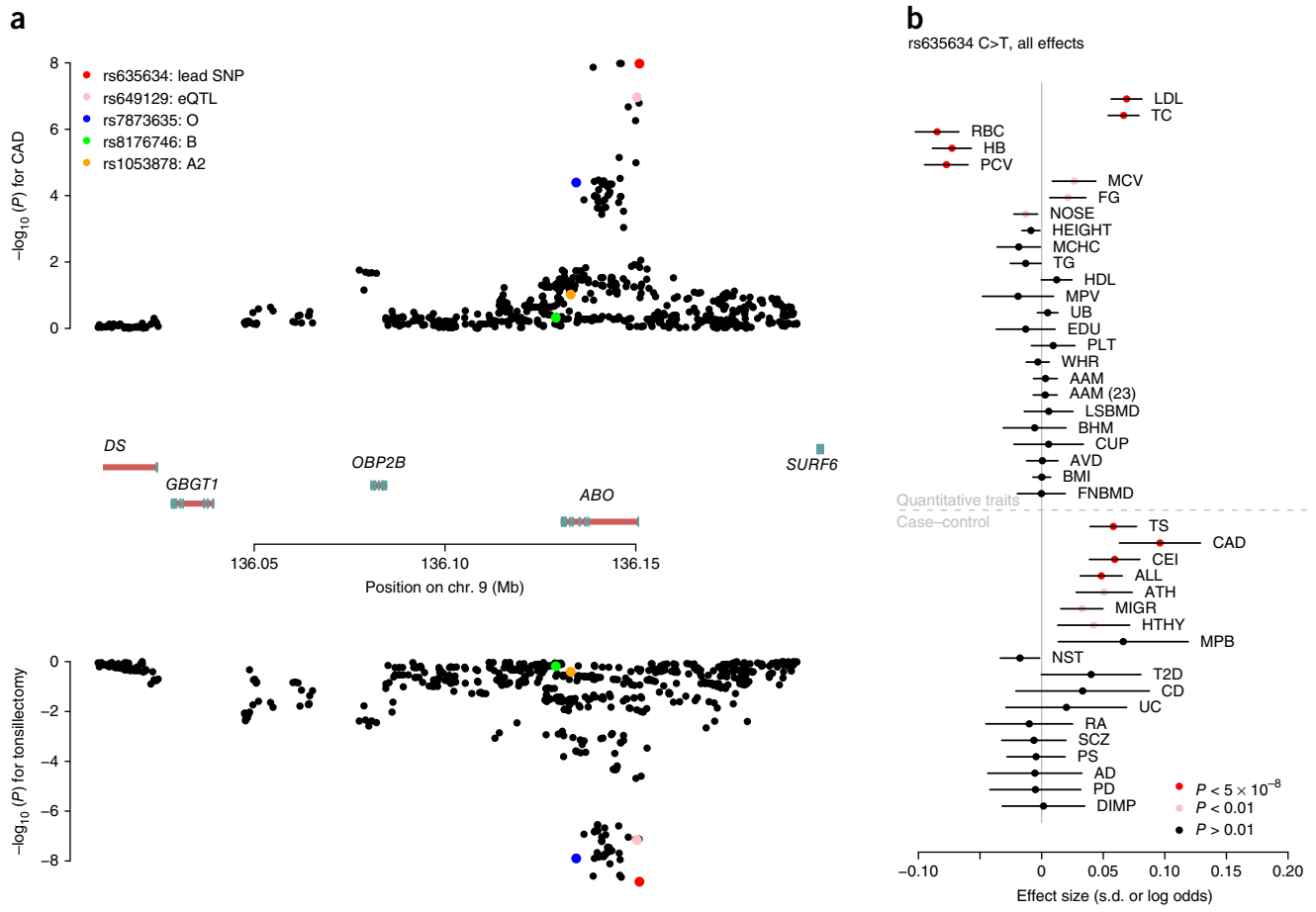

Figure 3 Multiple associations near the *ABO* gene. (**a**) Association signals for CAD and tonsillectomy. In the top plot, we show the *P* values for association with CAD for variants in the window around the *ABO* gene. In the bottom plot are the *P* values for association with tonsillectomy. In both plots, the points representing SNPs tagging functionally important alleles at *ABO* are colored. Between the plots are the gene models for the region; exons are denoted by blue boxes, and introns are denoted by red lines. Note that the *ABO* gene is transcribed from the negative strand. (**b**) Association effect sizes for rs635634 on all tested traits. Shown are the effect size estimates for rs635634 for all traits. The lines represent 95% confidence intervals. Traits are grouped according to whether they are quantitative traits (in which case, the *x* axis is in units of s.d.) or case–control traits (in which case, the *x* axis is in units of log OR).

Finally, we evaluated the method in simulations (**Supplementary Figs. 1**–**5**) and found that the model gives a small overestimate of the proportion of shared effects (**Supplementary Fig. 3**). This is because the amount of evidence against the null model of no associations is greater when a variant influences both phenotypes as compared to when it only influences a single phenotype (**Supplementary Fig. 4**).

Overlapping association signals identified in 43 GWAS

We applied the method to all pairs of the 43 GWAS listed in **Table 1**. For each pair of studies, we first estimated the expected correlation in the effect sizes from the summary statistics and included this correction for overlapping individuals in the model. Note that this is conservative: in pairs of GWAS where we are sure that there are no overlapping individuals (for example, age at menarche and age at voice drop), we saw that the correlation in the summary statistics was nonzero, indicating that we are correcting out some truly shared genetic effects on the two traits (**Supplementary Fig. 6**).

To gain an exploratory sense of the relationships between the phenotypes, we examined the patterns of overlap in associations among all 43 studies. Specifically, the model can be used to estimate, for each pair of traits [*i*,*j*], the proportion of detected variants that influence trait *i* that also detectably influence trait *j*. These estimates are shown in **Figure 2**, with phenotypes clustered according to their patterns of overlap. We see several clusters of related traits. For example, of the variants that detectably influence age at menarche (in the study by Perry *et al.*[43]), the maximum a posteriori estimate is that 36% detectably influence height, 30% detectably influence age at voice drop, 28% influence BMI, 10% influence breast size, and 10% influence male-pattern baldness. We interpret this as a set of phenotypes that share hormonal regulation. Additionally, there is a large cluster of phenotypes including coronary artery disease (CAD), type 2 diabetes, red blood cell traits, and lipid traits, which we interpret as a set of metabolic traits. Further, immune-related disease (allergies, asthma, hypothyroidism, Crohn's disease, and rheumatoid arthritis) all cluster together and also cluster with infectious disease traits (childhood ear infections and tonsillectomy). This biologically relevant clustering validates the principle that GWAS variants can identify shared mechanisms underlying pairs of traits in a systematic way. As a control, we performed the same clustering of phenotypes by the estimated proportion of genomic regions where two causal sites fall nearby (model 4 in **Fig. 1**). In this case, there was no biologically meaningful clustering (**Supplementary Fig. 7**).

Individual loci that influence many traits

We next examined the individual loci identified by these pairwise GWAS. We identified 341 genomic regions where we infer the presence of a variant that influences a pair of traits, at a threshold of a posterior probability greater than 0.9 of model 3 (**Supplementary Table 1**). This number excludes 'trivial' findings where a genetic variant influences two similar traits (two lipid traits, two red blood cell traits, two platelet traits, both measures of BMD, both inflammatory bowel diseases, or type 2 diabetes and fasting glucose) and the MHC region. A previous 'phenome-wide association study' identified 44 genetic variants associated with multiple phenotypes[34], so this represents an order of magnitude increase in the number of such loci.

Some genomic regions contain variants that influence a large number of the traits we considered. We ranked each genomic region according to how many phenotypes share genetic associations in the region (that is, if the pairwise scan for both height and CAD and the pairwise scan for CAD and LDL both indicated the same region, we counted this as three phenotypes sharing an association in the region). The top region in this ranking identified a nonsynonymous polymorphism in *SH2B3* (rs3184504) that is associated with a number of autoimmune diseases, lipid traits, heart disease, and red blood cell traits (**Supplementary Fig. 8** and **Supplementary Table 2**). This variant has been identified in many GWAS, particularly for autoimmune diseases[44].

The next region in this ranking contains the gene encoding the ABO histo-blood groups in humans and has a variant associated with 11 traits in these data (and many other additional traits not in these data; see also refs. 20,45–47). In **Figure 3a**, we show the association statistics in this region for CAD and probability of having a tonsillectomy. At the lead SNP, the non-reference allele is associated with increased risk of CAD ($z = 5.7$, $P = 1.1 \times 10^{-8}$) and increased risk of having a tonsillectomy ($z = 6.0$, $P = 1.5 \times 10^{-9}$). This variant is also strongly associated with other immune, red blood cell, and lipid traits in these data (**Fig. 3b**). A tag for a microsatellite that influences the expression of *ABO*[48] is correlated with the lead SNP rs635634, as is a tag for the O blood group (**Fig. 3a**). However, the lead SNP is an expression quantitative trait locus (eQTL) for both *ABO* and the nearby gene *SLC2A6* in whole blood[46], so this allele may in fact have downstream effects via effects on the expression of two genes.

Among the top ranked regions were several where the likely causal variant is known: (i) a nonsynonymous variant in the zinc transporter *SLC39A8* (rs13107325; **Supplementary Fig. 9**) that is associated with schizophrenia (log OR for the non-reference allele = 0.15, $P = 2 \times 10^{-12}$), Parkinson disease (log OR = −0.15, $P = 1.6 \times 10^{-7}$), and height ($\hat{\beta}$ = −0.03 s.d., $P = 3.8 \times 10^{-7}$), among others; (ii) a nonsynonymous variant in the glucokinase regulator *GCKR* (rs1260326; **Supplementary Fig. 10**) that is associated with fasting glucose levels ($\hat{\beta}$ = 0.06 s.d., $P = 5 \times 10^{-25}$) and height ($\hat{\beta}$ = 0.019 s.d., $P = 2.6 \times 10^{-11}$), among others; (iii) a set of variants near the *APOE* gene (which we presume to be driven by the APOE4 allele; **Supplementary Fig. 11**) that is associated with nearsightedness (rs6857: log OR = −0.04, $P = 1.8 \times 10^{-5}$), waist–hip ratio ($\hat{\beta}$ = −0.02 s.d., $P = 8.3 \times 10^{-5}$), and several lipid traits apart from the well-known association with Alzheimer disease; and (iv) regulatory variants in an intron of the *FTO* gene[49,50] that are associated with breast size in women (rs1421085: $\hat{\beta}$ = 0.06 s.d., $P = 3.5 \times 10^{-7}$; **Supplementary Fig. 12**) and age at voice drop in men ($\hat{\beta}$ = −0.02 s.d., $P = 2.7 \times 10^{-5}$), among others.

It has previously been observed that association signals for different phenotypes tend to cluster spatially in the genome[51]; these results suggest that, in some cases, clustered associations are driven by single variants. We note anecdotally that the variants that influence a large number of phenotypes often seem to be nonsynonymous rather than regulatory changes, which contrasts with the pattern seen in association studies overall (for example, see ref. 37).

Identifying pairs of phenotypes with correlated effect sizes

In our scan for variants that influence pairs of phenotypes, we did not assume any relationship between the effect sizes of a variant on the two phenotypes. However, if two traits are influenced by shared underlying molecular mechanisms, we might expect the effects of a variant on the two phenotypes to be correlated. To test this hypothesis, we returned to the set of variants identified by analysis of each phenotype individually (the numbers of these variants for each trait are given in **Table 1**). For each set, we calculated the rank correlation

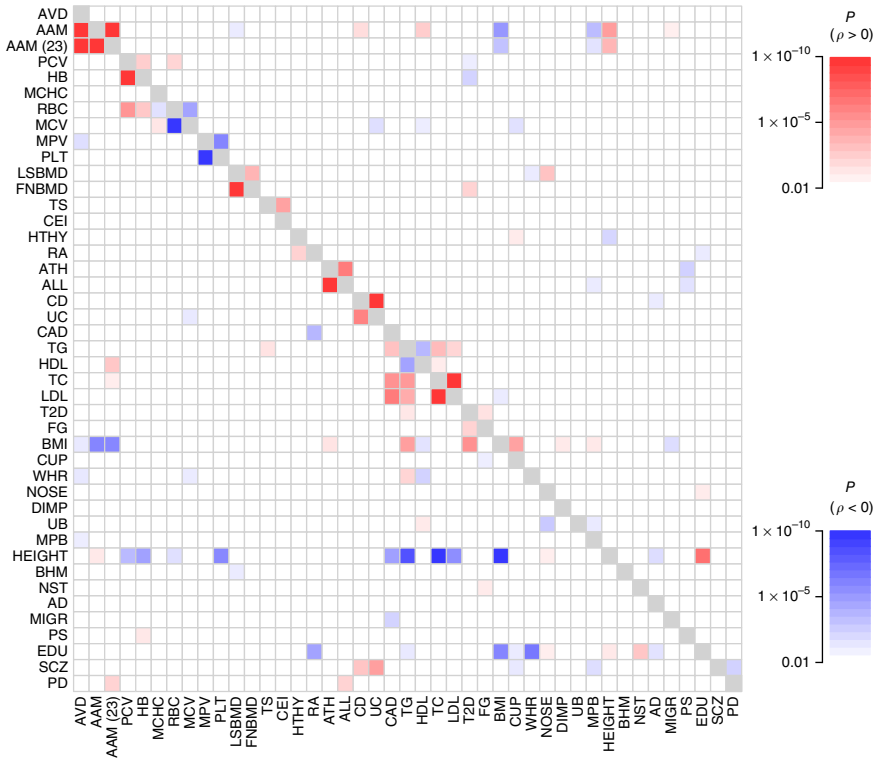

Figure 4 Heat map showing patterns of correlated effect sizes for variants across pairs of traits. For each pair of traits [*i*,*j*], we extracted the set of variants that influence trait *i* and their effect sizes on both *i* and *j*. We then calculated Spearman's rank correlation between the effect sizes on *i* and the effect sizes on *j* and tested whether this correlation was significantly different from zero. Shown in color are all pairs of traits where this test gave $P < 0.01$. Darker colors correspond to smaller *P* values, and color corresponds to the direction of the correlation (red, positive; blue, negative). The phenotypes are in the same order as in **Figure 2**. For a comparison to genome-wide genetic correlations, see **Supplementary Figure 13**.

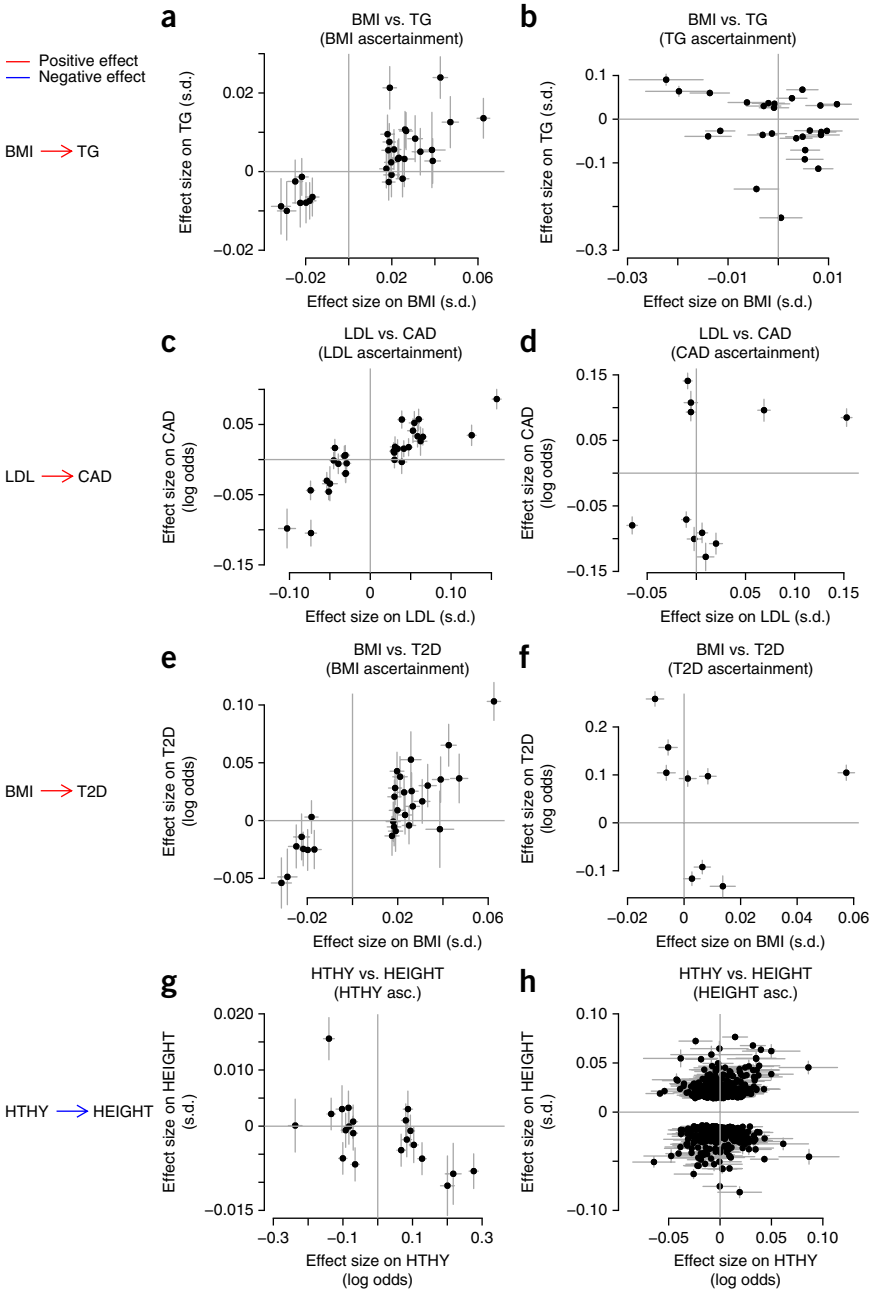

Figure 5 Putative causal relationships between pairs of traits. For each pair of traits identified as candidates to be related in a causal manner (Online Methods), we show the effect sizes of genetic variants on the two traits (at genetic variants successfully genotyped or imputed in both studies). Lines represent one standard error. (**a**,**b**) BMI and triglyceride levels. The effect sizes of genetic variants on BMI and triglyceride levels are shown for variants identified in the GWAS for BMI (**a**) or triglycerides (**b**). (**c**,**d**) LDL and CAD. The effect sizes of genetic variants on LDL levels and CAD are shown for variants identified in the GWAS for LDL (**c**) or CAD (**d**). (**e**,**f**) BMI and type 2 diabetes. The effect sizes of genetic variants on BMI and type 2 diabetes are shown for variants identified in the GWAS for BMI (**e**) or type 2 diabetes (**f**). (**g**,**h**) Hypothyroidism and height. The effect sizes of genetic variants on hypothyroidism and height are shown for variants identified in the GWAS for hypothyroidism (**g**) or height (**h**).

between the effect sizes of the variants on the index trait (the one in which the variants were identified) and all of the other traits.

The results of this analysis are presented in **Figure 4**. Apart from closely related traits (for example, the two measurements of bone density), we saw a number of traits that were correlated at a genetic level. We focus on two of these. First, variants associated with delayed age of menarche in women tend, on average, to be associated with decreased BMI ($\rho = -0.53$, $P = 1.2 \times 10^{-6}$), reduced risk of male-pattern baldness ($\rho = -0.45$, $P = 5.9 \times 10^{-5}$), and increased height ($\rho = 0.52$, $P = 2.2 \times 10^{-6}$; **Fig. 4**). These patterns held both for the GWAS on age at menarche performed by Perry *et al.*[43] and that performed by 23andMe (**Fig. 4**). Most of these variants also delay age at voice drop in men (**Fig. 2**), so we interpret these variants as ones that influence pubertal timing in general. The negative correlation between a variant's effect on age at menarche and BMI has previously been observed[39,43,52], as has the positive correlation between a variant's effect on age at menarche and height[39,43]. The negative correlation between a variant's effect on age at menarche (or, more likely, puberty in general) and male-pattern baldness has not been previously noted but is consistent with the known role for increased androgen signaling in causing hair loss[53–55].

Second, we found that genetic variants associated with increased risk of schizophrenia tended to be associated with increased risk of both Crohn's disease ($\rho = 0.27$, $P = 2.2 \times 10^{-4}$) and ulcerative colitis ($\rho = 0.33$, $P = 6.6 \times 10^{-6}$). These correlations (identified only at the most strongly associated SNPs) are also present at the level of genome-wide genetic correlations between the diseases[39] (**Supplementary Fig. 13**). This observation is consistent with slightly higher rates of autoimmune diseases (including Crohn's disease and ulcerative colitis) in patients with schizophrenia in Denmark[56–58] and with molecular evidence for a partial autoimmune etiology for schizophrenia (for example, see ref. 59).

Inferring causal relationships between traits

Finally, we were interested in identifying pairs of traits that may be related in a causal manner. Because we are using observational data (rather than, for example, a randomized controlled trial), we view strong statements about causality as impossible. Nonetheless, a realistic goal might be to identify aspects of the data that are more consistent with a causal model than a non-causal model.

As a motivating example, we considered the correlation between levels of LDL cholesterol and risk of CAD, now widely accepted as a causal relationship[60]. We noticed that variants ascertained as having an effect on LDL cholesterol levels had correlated effects on risk of CAD (**Figs. 4** and **5c**), whereas variants ascertained as having an effect on CAD risk did not in general have correlated effects on LDL levels (**Fig. 5d**). This is consistent with the hypothesis that LDL cholesterol is one of many causal factors that influence CAD risk. An alternative interpretation is that LDL

cholesterol is highly genetically correlated to an unobserved trait that causally influences risk of CAD.

We developed a method to detect pairs of traits that show this asymmetry in the effect sizes of associated variants, which we interpret as more consistent with a causal relationship between the traits than a non-causal one (Online Methods). At a threshold of a relative likelihood of 100 in favor of a causal versus a non-causal model, we identified five pairs of putative causally related traits. (At a less stringent threshold of a relative likelihood of 20 in favor of a causal model, we identified 11 additional pairs of traits (**Supplementary Fig. 14**).) Simulations suggest that this threshold corresponds approximately to a *P* value around 0.001 (**Supplementary Fig. 15**) and that the power of this test depends on the number of genetic variants used as input and the true underlying correlation in their effect sizes (**Supplementary Fig. 16**). Four of these are shown in **Figure 5**. First, genetic variants that influence BMI had correlated effects on triglyceride levels, whereas the reverse was not true; this suggests that increased BMI is a cause for increased triglyceride levels (**Fig. 5**). Randomized controlled trials of weight loss are also consistent with this causal link[61,62], as are Mendelian randomization studies[63,64]. Second, we confirmed the evidence in favor of a causal role for increased LDL cholesterol levels in CAD (**Fig. 5**) and in favor of a causal role for increased BMI in type 2 diabetes risk (**Fig. 5** and **Supplementary Fig. 17**). Finally, we suggest that increased risk of hypothyroidism causes decreased height (**Fig. 5**). Although it is known that severe hypothyroidism in childhood leads to decreased adult height (for example, see ref. 65), these data indicated that hypothyroidism susceptibility may also influence height in the general population. A fifth potentially causal relationship (between risk of CAD and rheumatoid arthritis) could not be confirmed in a larger study and so is not displayed (**Supplementary Fig. 18** and **Supplementary Note**).

DISCUSSION

We have performed a scan for genetic variants that influence multiple phenotypes and have identified several hundred loci that influence multiple traits. This style of scan complements methods to quantify the genetic correlation between two traits[39,41,66,67], which are not generally concerned with identifying individual variants that influence both traits. We were interested in using the individual variants found to affect multiple traits to identify biological relationships between traits, including potential relationships where one trait is causally upstream of the other. Other potential mechanisms that could lead to an association between a genetic variant and two phenotypes include transgenerational effects for a variant, with one effect on a parental phenotype and an effect on a separate phenotype in the offspring (for example, see refs. 68,69), or assortative mating that involves more than one trait[70].

A number of limitations of this study are worth mentioning. First, all of the GWAS we have used are based on genotyping arrays and imputation, and thus the loci identified are generally common (minor allele frequency over 1%). Inferences from common variants such as these may not hold for rarer variants that may emerge from large sequencing studies. Second, we reiterate that all of our inferences are based on sets of 'detectable' loci; the GWAS we have used have highly variable sample sizes, and the traits have variable genetic architectures. As sample sizes for all traits reach the millions, inferences from detectable loci will converge to inferences from all loci. If traits truly follow an infinitesimal model (where every genetic variant influences every trait), we speculate that patterns of genetic overlap (such as those in **Fig. 2**) will become less interpretable, while patterns of genetic correlation (such as those in **Fig. 4**) may be more useful.

One clear observation from these data is that genetic variants that influence puberty (age at menarche and age at voice drop) often have correlated effects on BMI, height, and male-pattern baldness (**Fig. 4**). In our scan for causal relationships between traits, we found modest evidence of a causal role of age at menarche in influencing adult height and for a causal role of BMI in the development of male-pattern baldness (**Supplementary Fig. 12**). The non-causal alternative (also consistent with the data) is that all of these traits are influenced by some of the same underlying biological pathways, and perhaps the most likely candidate for this pathway is hormonal signaling. This highlights the importance of considering evidence from multiple traits when interpreting the molecular consequences of a variant and designing experimental studies. Although variants that influence height overall are enriched near genes expressed in cartilage[71] and variants that influence BMI are enriched near genes expressed broadly in the central nervous system[72], it seems that a subset of these variants also influence age at menarche and male-pattern baldness. For these variants, it may be worth considering functional follow-up in gonadal tissues or specific brain regions known to be important in hormonal signaling.

It is also striking to note how many genetic variants influence multiple traits (**Fig. 2**) but without a consistent correlation in effect sizes (**Fig. 4**). For example, many of the autoimmune and immune-related traits appear to have many genetic causes in common, but the effect sizes of the variants on the different traits seem to be largely uncorrelated (see also refs. 10,39). Likewise, many variants appear to influence lipid traits, red blood cell traits, and immune traits, but without consistent directions of effect. A trivial explanation for this observation is that we are underpowered to detect correlations in effect sizes because we are using only a small set of the SNPs with the strongest associations. However, the genetic correlations between many of these traits (calculated using all SNPs) are not significantly different from zero[39] (**Supplementary Fig. 13**). Another possibility is that a given genetic variant often influences the function of multiple cell types through separate molecular pathways or that the effects of a variant on two related phenotypes vary according to an individual's environmental exposures.

From the point of view of epidemiology, the ability to scan through many pairs of traits to find those that are potentially causally related seems appealing, and some previous analyses have had similar goals[73]. Our approach makes the key assumption that, if two traits are related in a causal manner, then the 'causal' trait is one of many factors that influence the 'caused' trait. This results in an asymmetry in the effects of genetic variants on the two traits that can be detected (**Fig. 5**). We also assume that we have identified a modest number of variants that influence both traits. This naturally means we are limited to considering heritable traits that have been studied within cohorts with moderate sample sizes (on the order of tens to hundreds of thousands of individuals). It seems likely that the main limiting factor to scaling this approach (should it be generally useful) will be phenotyping rather than genotyping.

URLs. gwas-pw code, https://github.com/joepickrell/gwas-pw; approximately independent LD blocks, https://bitbucket.org/nygcresearch/ldetect-data.

METHODS

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

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

ACKNOWLEDGMENTS

This work was supported in part by the National Human Genome Research Institute of the National Institutes of Health (grant R44HG006981 to 23andMe) and the National Institute of Mental Health (grant R01MH106842 to J.K.P.). We thank the customers of 23andMe for making this work possible, the GWAS consortia that made summary statistics available to us, L. Jostins for providing updated summary statistics from the Crohn's disease and ulcerative colitis GWAS, and G. Coop and M. Stephens for helpful discussions. We thank D. Golan and J. Pritchard for comments on a previous version of this manuscript. We thank D. Cesarini and the Social Science Genetic Association Consortium for access to summary statistics from the association study of educational attainment.

Data on glycemic traits have been contributed by MAGIC investigators and have been downloaded from http://www.magicinvestigators.org/. Data on CAD and myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from http://www.cardiogramplusc4d.org/.

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The iSelect chips were funded by the French National Foundation on Alzheimer disease and related disorders. EADI was supported by LABEX (Laboratory of Excellence program investment for the future) DISTALZ grant, INSERM, Institut Pasteur de Lille, Université de Lille 2, and the Lille University Hospital. GERAD was supported by the Medical Research Council (grant 503480), Alzheimer's Research UK (grant 503176), the Wellcome Trust (grant 082604/2/07/Z), and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grants 01GI0102, 01GI0711, and 01GI0420. CHARGE was partly supported by NIH/NIA grant R01 AG033193 and NIA grant AG081220 and AGES contract N01-AG-12100, NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by NIH/NIA grants U01 AG032984, U24 AG021886, and U01 AG016976, and by Alzheimer's Association grant ADGC-10-196728.

AUTHOR CONTRIBUTIONS

J.K.P. developed and applied the methods for pairwise analysis of association studies. T.B. contributed to the splitting of GWAS hits into independent blocks. J.Z.L. performed the LD score regression analyses. L.S. contributed to the analysis of the *ABO* region. J.Y.T. and D.A.H. performed and analyzed the studies from 23andMe. All authors contributed to the writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Stearns, F.W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–773 (2010).
2. Paaby, A.B. & Rockman, M.V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).
3. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
4. Chillón, M. *et al.* Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *N. Engl. J. Med.* **332**, 1475–1480 (1995).
5. Müller, C. Xanthomata, hypercholesterolemia, angina pectoris. *Acta Med. Scand.* **95**, 75–84 (1938).
6. Steinberg, D. Atherogenesis in perspective: hypercholesterolemia and inflammation as partners in crime. *Nat. Med.* **8**, 1211–1217 (2002).
7. Pearl, J. *Causality: Models, Reasoning and Inference* vol. 29 (Cambridge University Press, 2000).
8. Steinberg, D. The cholesterol controversy is over. Why did it take so long? *Circulation* **80**, 1070–1078 (1989).
9. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
10. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).
11. Andreassen, O.A. *et al.* Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* **92**, 197–209 (2013).
12. Andreassen, O.A. *et al.* Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet.* **9**, e1003455 (2013).
13. Elliott, K.S. *et al.* Evaluation of the genetic overlap between osteoarthritis with body mass index and height using genome-wide association scan data. *Ann. Rheum. Dis.* **72**, 935–941 (2013).
14. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
15. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
16. Styrkarsdottir, U. *et al.* Nonsense mutation in the *LGR4* gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
17. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501 (2012).
18. Moltke, I. *et al.* A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
19. Pendergrass, S.A. *et al.* Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet.* **9**, e1003087 (2013).
20. Li, L. *et al.* Disease risk factors identified through shared genetic architecture and electronic medical records. *Sci. Transl. Med.* **6**, 234ra57 (2014).
21. Katan, M.B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* **1**, 507–508 (1986).
22. Smith, G.D. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidemiol.* **33**, 30–42 (2004).
23. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23** R1, R89–R98 (2014).
24. Voight, B.F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
25. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).
26. Panoutsopoulou, K. *et al.* The effect of *FTO* variation on increased osteoarthritis risk is mediated through body mass index: a Mendelian randomisation study. *Ann. Rheum. Dis.* **73**, 2082–2086 (2014).
27. Holmes, M.V. *et al.* Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *Am. J. Hum. Genet.* **94**, 198–208 (2014).
28. De Silva, N.M.G. *et al.* Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes* **60**, 1008–1018 (2011).
29. Granell, R. *et al.* Effects of BMI, fat mass, and lean mass on asthma in childhood: a Mendelian randomization study. *PLoS Med.* **11**, e1001669 (2014).
30. Pichler, I. *et al.* Serum iron levels and the risk of Parkinson disease: a Mendelian randomization study. *PLoS Med.* **10**, e1001462 (2013).
31. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* **14**, 661–673 (2013).
32. Fortune, M.D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
33. 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* **381**, 1371–1379 (2013).
34. Denny, J.C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
35. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
36. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
37. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
38. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
39. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
40. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
41. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
42. Flint, J. & Mackay, T.F.C. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.* **19**, 723–733 (2009).
43. Perry, J.R.B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
44. Richard-Miceli, C. & Criswell, L.A. Emerging patterns of genetic overlap across autoimmune disorders. *Genome Med.* **4**, 6 (2012).
45. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
46. Wessel, J. *et al.* Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* **6**, 5897 (2015).
47. Franchini, M. & Lippi, G. The intriguing relationship between the *ABO* blood group, cardiovascular disease, and cancer. *BMC Med.* **13**, 7 (2015).
48. Kominato, Y., Tsuchiya, T., Hata, N., Takizawa, H. & Yamamoto, F. Transcription of human *ABO* histo-blood group genes is dependent upon binding of transcription factor CBF/NF-Y to minisatellite sequence. *J. Biol. Chem.* **272**, 25890–25898 (1997).

49. Smemo, S. *et al.* Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).
50. Claussnitzer, M. *et al. FTO* obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
51. Jeck, W.R., Siebold, A.P. & Sharpless, N.E. Review: a meta-analysis of GWAS and age-associated diseases. *Aging Cell* **11**, 727–731 (2012).
52. Elks, C.E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.* **42**, 1077–1085 (2010).
53. Li, R. *et al.* Six novel susceptibility loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet.* **8**, e1002746 (2012).
54. Richards, J.B. *et al.* Male-pattern baldness susceptibility locus at 20p11. *Nat. Genet.* **40**, 1282–1284 (2008).
55. Hamilton, J.B. Patterned loss of hair in man; types and incidence. *Ann. NY Acad. Sci.* **53**, 708–728 (1951).
56. Eaton, W.W. *et al.* Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *Am. J. Psychiatry* **163**, 521–528 (2006).
57. Eaton, W. *et al.* Coeliac disease and schizophrenia: population based case control study with linkage of Danish national registers. *Br. Med. J.* **328**, 438–439 (2004).
58. Benros, M.E. *et al.* Autoimmune diseases and severe infections as risk factors for schizophrenia: a 30-year population-based register study. *Am. J. Psychiatry* **168**, 1303–1310 (2011).
59. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
60. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* **344**, 1383–1389 (1994).
61. Pi-Sunyer, X. *et al.* Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the Look AHEAD trial. *Diabetes Care* **30**, 1374–1383 (2007).
62. Shai, I. *et al.* Weight loss with a low-carbohydrate, Mediterranean, or low-fat diet. *N. Engl. J. Med.* **359**, 229–241 (2008).
63. Würtz, P. *et al.* Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med.* **11**, e1001765 (2014).
64. Freathy, R.M. *et al.* Common variation in the *FTO* gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes* **57**, 1419–1426 (2008).
65. Rivkees, S.A., Bode, H.H. & Crawford, J.D. Long-term growth in juvenile acquired hypothyroidism: the failure to achieve normal adult stature. *N. Engl. J. Med.* **318**, 599–602 (1988).
66. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
67. Visscher, P.M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
68. Ueland, P.M., Hustad, S., Schneede, J., Refsum, H. & Vollset, S.E. Biological and clinical implications of the MTHFR C677T polymorphism. *Trends Pharmacol. Sci.* **22**, 195–201 (2001).
69. Zhang, G. *et al.* Assessing the causal relationship of maternal height on birth size and gestational age at birth: a Mendelian randomization analysis. *PLoS Med.* **12**, e1001865 (2015).
70. Gianola, D. Assortative mating and the genetic correlation. *Theor. Appl. Genet.* **62**, 225–231 (1982).
71. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
72. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
73. Evans, D.M. *et al.* Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genet.* **9**, e1003919 (2013).
74. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).
75. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
76. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* http://dx.doi.org/10.1038/nature17671 (2016).
77. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
78. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
79. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
80. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
81. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
82. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
83. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
84. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).

ONLINE METHODS

Study overview. The sources of the GWAS data analyzed in this study are described in detail in the **Supplementary Note**. For each study, we imputed summary statistics or genotypes for all autosomal variants in the March 2012 release of the 1000 Genomes Project Phase 1 (ref. 35). Our method uses the *z* scores and standard errors of the estimated effect sizes for each SNP. In studies where standard errors were not provided, we approximated them using the allele frequencies from the European-descent individuals in the 1000 Genomes Project Phase 1 release and the reported sample size of the study (see ref. 37). Throughout the paper, we report effect sizes of variants as the effect of the non-reference allele in human genome reference hg19.

Hierarchical model. The hierarchical model used for the main scan for overlapping association signals in two GWAS data sets is described in detail in the **Supplementary Note**. Software implementing the model is available through GitHub (see URLs).

Causal inference. We aimed to develop a robust method for measuring the evidence in favor of a causal relationship between two traits using data from many genetic associations, while recognizing that strong conclusions are likely impossible in this setting. The approach we developed is described in detail in the **Supplementary Note**.

doi:10.1038/ng.3570