Identifying the genetic basis and molecular mechanisms underlying phenotypic correlation between complex human traits using a gene-based approach

Jialiang Gu[1*], Chris Fuller[1,2*], Jiashun Zheng[1#], Hao Li[1#]
1. Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158, USA,
2. Caribou Biosciences, 2929 Seventh Street #105, Berkeley, CA 94710
* Both authors contributed equally to this work.
# send correspondence to haoli@genome.ucsf.edu and jiashun@genome.ucsf.edu

Abstract

Phenotypic correlations between complex human traits have long been observed based on epidemiological studies. However, the genetic basis and underlying mechanisms are largely unknown. The recent accumulation of GWAS data has made it possible to analyze the genetic similarity between human traits through comparative analysis. Here we developed a gene-based approach to measure genetic similarity between a pair of traits and to delineate the shared genes/pathways, through three steps: 1) translating SNP-phenotype association profile to gene-phenotype association profile by integrating GWAS with eQTL data; 2) measuring the similarity between a pair of traits by a normalized distance between the two gene-phenotype association profiles; 3) delineating genes/pathways supporting the similarity. Application of this approach to a set of GWAS data covering 59 human traits detected significant similarity between many known and unexpected pairs of traits; a significant fraction of them are not detectable by SNP based similarity measures. Examples include Height and Schizophrenia, Cancer and Alzheimer's Disease, and Rheumatoid Arthritis and Crohn's disease. Functional analysis revealed specific genes/pathways shared by these pairs. For example, Height and Schizophrenia are co-associated with genes involved in neural development, skeletal muscle regeneration, protein synthesis, magnesium homeostasis, and immune response, suggesting growth and development as a common theme underlying both traits. Our approach can detect yet unknown relationships between complex traits and generate mechanistic hypotheses, and has the potential to improve diagnosis and treatment by transferring knowledge from one disease to another.

Introduction

The human body is a complex system with multiple interacting components, and as such many traits are inter-related. Phenotypic correlations between complex traits have long been observed based on epidemiological studies. Depending on the level of our understanding of the underlying mechanism, some of the relations may seem obvious, e.g., Fasting Insulin vs. Fasting Glucose, while others may seem totally unexpected, such as the inverse correlation between Height and risk for Schizophrenia[1]. It has been argued that many of the phenotypic correlations have genetic underpinning, based on the inheritance patterns of the traits involved. However, little is known due to the polygenic nature of these traits. Understanding phenotypical connections at the genetic level can potentially lead to new mechanistic insight and allow the transfer of diagnosis and treatment from one well-studied disease to another.

The recent application of GWAS to a wide range of complex human traits has provided an opportunity to systematically identify genetic similarity between traits and delineate the underlying mechanism. Several comparative GWAS analysis tools have been developed to detect genetic similarity between phenotypes based on co-associations at the SNP level, and applications of these tools to GWASs have lent genetic support to some of the earlier epidemiological observations[2-7]. However, the genetic similarity between different traits can be obscured at the SNP level. Phenotypes sharing a set of intermediate molecular traits such as gene expression may not necessarily associate with the same SNPs, since multiple SNPs can converge on the same gene, and only a small subset of these SNPs may be identified in a particular GWAS. Furthermore, it is challenging to infer common molecular mechanisms based on the shared SNPs, as they often fall into non-coding regions with no apparent functional implication. Therefore, It is logical to consider genetic similarity at the gene level that integrates information from multiple convergent SNPs.

Here we introduce a gene-based approach to detect genetic similarity between traits, and to identify shared genes/pathways. The first step in this approach is to assign a p-value of association for each gene with a given phenotype, so that each phenotype is represented by a gene association vector of dimension N_g, while N_g is the number of genes in the genome. This step translates the SNP-phenotype association profile from the original GWAS into a gene-phenotype association profile. For this purpose, we developed a new computational algorithm called Sherlock-II, to integrate GWAS with eQTL data using the collective information of all SNPs. In the second step, we score the similarity between two phenotypes based on their gene-phenotype association vectors, using an appropriately normalized distance between the two. We then assess the statistical significance of the measured similarity score using a background distribution generated by comparing the real GWAS data with "fake" GWASs generated from randomly selected cases and controls. Once the similarity between two phenotypes is detected, we identify groups of functionally related genes (such as those defined by gene ontology (GO) terms or KEGG pathways) that contribute to the similarity using the Partial Pearson Correlation Analysis (PPCA) that we previously developed [8]. This generates hypotheses regarding common physiological processes underlying different and sometimes seemingly unrelated phenotypes.

Using this approach, we analyzed a dataset of ~400 GWASs, using a combined eQTL dataset, including GTEx[9]. We measured pairwise similarity between 59 phenotypes and identified many known and unexpected relationships. Many of the discovered relationships are supported by epidemiological observations. Examples include Height and Schizophrenia, Alzheimer's Disease (AD) and cancer, and Rheumatoid Arthritis and Crohn's disease. Functional analysis revealed a broad spectrum of genes shared by Height and Schizophrenia, with a common theme on growth and development. Our approach can detect genetic similarity between seemingly unrelated phenotypes that are not detectable by SNP-based similarity measures[5], and importantly, for each of the detected pairs, our method suggests a number of shared genes/pathways that point to a potential mechanistic connection between the two phenotypes.

Results

Translate SNP-phenotype association profile to gene-phenotype association profile using Sherlock-II algorithm

The first step in our gene-based approach is to assign a p-value of association for each gene to a given phenotype, thus translating the SNP association profile of the original GWAS to a gene association profile. To have better power in detecting genetic similarity, such transformation should preserve as much information from the GWAS as possible (not just the SNPs passing the genome-wide significance threshold) and utilize the convergent effect of multiple SNPs on the same gene. Thus, we need a high throughput tool to combine GWAS data with eQTL data that can capture the collective information from multiple SNPs in relation to a gene, both in cis and trans. In addition, due to the scale of computation, this tool should be statistically robust and computationally efficient.

Multiple computational tools have been developed to associate complex traits with genes [10-16]. However, most of the existing tools infer the association between gene and trait by colocalizing the GWAS and eQTL signal in a single locus, starting from a strong GWAS or eQTL peak. Such an approach leads to predictions with high specificity but may miss genes supported by the alignment of multiple GWAS and eQTL peaks that do not pass the genome-wide significance threshold. Previously, He et al. developed the Sherlock algorithm that use the collective information of all the SNPs to infer gene-phenotype association[13]. However, the Bayesian formulation used makes it difficult to compute p-values and is sensitive to inflation in the input data. Here we developed a new computational algorithm called Sherlock-II, a second generation of Sherlock that overcomes the limitations of Sherlock, to integrate GWAS with eQTL data.

Similar to Sherlock, Sherlock-II detects gene-phenotype associations by comparing the GWAS SNPs with the eQTL SNPs of a gene, with the assumption that if the gene is causal to the phenotype, SNPs that influence the target gene's expression should also influence the phenotype; thus the eQTL profile of the gene should have significant overlap with the GWAS profile of a phenotype. Although conceptually similar, Sherlock-II uses a totally different statistical approach to calculate the statistical significance of the profile alignment, leading to a number of significant

improvements. While Sherlock calculates a Bayesian factor for the observed eQTL and GWAS profiles and uses randomization of GWAS data to estimate p-value, Sherlock-II uses a statistic that is a sum of the log(p-value) of the GWAS peaks aligned to the eQTL peaks, with background distribution of the statistic directly calculated by the convolution of the distribution of the log(p-value) for all the independent GWAS peaks (see Methods). This new approach gives Sherlock-II several advantages: 1) it is more robust against inflation. By deriving the background distribution empirically from the list of all p-values for the GWAS SNPs, it automatically accounts for inflation that may exist in the original GWAS data; 2) it gives more accurate and efficient p-value calculation since the background distribution can be calculated through convolution, instead of GWAS randomization; 3) Sherlock-II is essentially parameter-free, while Sherlock requires a number of prior parameters. These combined features of Sherlock-II made it possible to perform a global analysis of GWASs vs. eQTL data efficiently.

Using Sherlock-II, we have analyzed a set of the published GWAS data, using the eQTL data we curated. The analyses include 445 GWAS datasets covering 88 phenotypes, 8 individual eQTL covering 6 unique tissues[17-24] and GTEx (version 7) data[25]. To include the possibility that the expression of a gene can influence the phenotype through multiple tissues, we combined eQTL data across different tissues: when a SNP appears in multiple eQTL datasets, we assign the most significant p-value to it to capture the best eQTL signal across tissues. This merged version includes the eQTL profile of 21,892 genes. For each GWAS dataset, Sherlock-II calculates a p-value for every gene measuring the statistical significance of the overlap between the merged eQTL profile of the gene and the GWAS profile, thus translating the original SNP-phenotype association profile to a gene-phenotype association profile. Overall, we have identified 1201 gene-phenotype associations with FDR <0.1, covering 504 genes and 56 phenotypes (Dataset S1). Many of these findings have literature support and are biologically meaningful.

In many cases, although only weak SNP associations are present in the GWAS data, Sherlock-II detects strong gene-phenotype associations as many GWAS SNPs are considered in aggregate during the transformation. These strong associations are often supported by multiple aligned SNPs with weak to moderate GWAS p values, and collectively they give a strong signal. Examples include the autism GWAS, of which the original signal is weak (no SNP passing genome-wide significance threshold)[26]. Sherlock-II identified 10 genes with FDR<0.1; four of them (ENPP4, MAPT, LEPR, MTHFR) have strong literature support[27-30] (Figure S1).

We have also applied Sherlock-II to metabolite-QTL data[31] and discovered 63 metabolite-phenotype associations at FDR<0.1, covering 48 metabolites and 19 phenotypes (Dataset S1). Sherlock-II identified several metabolites with a strong association with neuro-degenerative/neuro-psychiatric diseases, although the signal in some of the original GWAS data is very weak (Figure S2).

Measure genetic similarity between two phenotypes using the global gene association profiles and identify shared genes/pathways

The next step in the gene-based approach is to score the similarity between two phenotypes based on their gene association profiles. Starting with the p-value of association between a gene i and a trait t: $P_i(t)$, we pick a vector representation for the association profile of the trait t, $V(t) =$

$[Z_1(t), Z_2(t), \ldots, Z_N(t)]$ where $Z_i(t)$ is the z score corresponding to the p-value $P_i(t)$, based on a two tailed test using standard normal distribution (see Methods), and N is the total number of genes in the genome. We use an unsigned z score because we do not consider the direction of the gene's effect on the phenotype (see Discussion). We then measure the Euclidian distance between the two phenotypes $E\big(V(t_1), V(t_2)\big)$. We expect that similar phenotypes will have certain components of the vectors with similar values, and thus will give rise to a smaller distance. We define a similarity score between the two traits as $S(t_1, t_2) = \big[\overline{E_{1,2^*}} - E\big(V(t_1), V(t_2)\big)\big] / \sigma_{E_{1,2^*}}$ where $\overline{E_{1,2^*}}$ and $\sigma_{E_{1,2^*}}$ are the mean and standard deviation of the Euclidean distance generated by randomly permuting the components of one of the two vectors (i.e., assuming that the two vectors are un-correlated). This distance normalization is essential as it takes out the contribution to the distance due to the features of each individual vectors, since we are interested in the similarity between the two vectors, not the vectors themselves. The statistical significance of the similarity score is calculated from a null distribution, which is generated by comparing "fake" GWAS data (with randomly selected cases and controls) with the GWAS data we analyze (see Methods for details). This similarity score can be used to search for traits that are similar to a given trait, much like sequence similarity search via alignment.

We verified that the similarity score and the ranking of the pairs is not sensitive to the actual vector representation and the distance metric we use. For example, using $-log_{10}(P_i(t))$ as the vector component or using Manhattan distance as a distance metric gives very similar results (see Figure S3, 4).

One major advantage of gene-based approach over SNP-based methods is that once the similarity is detected, the information on genes/pathways that support the similarity can be readily delineated. A straightforward approach is to select the top genes for each trait with a p-value cutoff and analyze the overlap. This approach does yield interesting results for some trait pairs (an example will be discussed below). However, for the majority of trait pairs that we detected as similar, the top associated genes for the two traits do not overlap. This suggests that the similarity is driven by a more diffused signal distributed to a broad set of genes. To identify such set of genes, we employ an approach we previously developed to identify functionally related genes supporting the similarity between two gene expression profiles[8]. This approach assesses the contribution of a subset of genes (e.g., defined by a GO category or a KEGG pathway) to the Pearson correlation and assigns a z-score to the subset (see Methods); we called it Partial-Pearson-Correlation Analysis (PPCA). This allows the ranking of the significance of different subsets of genes based on their z-score (PPCA-zscore). This approach enables us to identify sets of GO terms and KEGG pathways that significantly contribute to the genetic similarity between a pair of traits and generate hypotheses regarding the mechanisms that connect the two traits.

Application of gene-based approach to 59 Human traits revealed many known and un-expected similarity and potential mechanisms

Using the approach described above, we analyzed the relationships between 59 Human traits covered by the 69 previously published GWAS data. This is a filtered version of the 445 GWAS datasets analyzed by Sherlock-II, with redundant data and many sub-phenotype groupings in the

same GWAS removed. For a few traits, we kept multiple GWAS studies published separately as positive controls.

Using the gene-association profiles generated by Sherlock-II, we calculated similarity scores for all 2346 pairs of the 69 GWASs and detected significant similarity for 407 pairs with p-value cutoff =1e-3 (expect 2 false positives). Here we present the top hits for a selected subset of 16 traits in Figure 1, including some highly prevalent diseases such as cancer, diabetes, AD, and coronary artery disease. The full result of the pairwise comparison is provided online (Dataset S2).

Many of the detected similarities are known or biologically meaningful based on our understanding of the relationship between the two traits. Besides obviously related pairs of traits, such as Diabetes vs. Fasting Glucose, Height vs. Body Mass Index (BMI), our method detected genetic similarity between traits known to share similar etiology. For example, Crohn's disease vs. Ulcerative Colitis, Schizophrenia vs. Bipolar Disorder, and Fasting Glucose vs. Coronary Artery Disease. Many of these relations were also detected previously through comparative GWAS analysis using a well-established SNP based similarity measure- LD score regression (LDSC) [5] (Figure 1, entries with a checkmark).

Our method also detected a number of pairs expected to share similar etiology but not detected by LDSC (Figure 1, entry marked *). One such example is the relation between RA and CD, known to share inflammatory/autoimmune origin (LDSC gives a regression coefficient r -0.029 and p-value 0.7255). Using PPCA of GO and KEGG terms (see Methods), we identified a number of gene groups shared by the two traits, with a strong theme in inflammation and immune responses, including MHC class II protein complexes and pyroptosis, which is a highly inflammatory form of programmed cell death, supporting the idea that cell death caused by dysregulated inflammatory response is a common etiology for the two diseases [32, 33].

Importantly, our gene-based approach identified a number of unexpected connections between seemingly unrelated traits (Figure 1, highlighted in red). Many of these relationships have support from previous epidemiological studies, but the genetic basis and mechanism for the connections are largely unknown. We discuss two examples in more details below.

The similarity between Height and Schizophrenia and the supporting genes/pathways

Both Schizophrenia and Height are highly inheritable, and clinical and epidemiological studies suggest that they are inversely correlated[1, 34-38]. Therefore, it is reasonable to assume that they share common genetic basis. However, a direct comparison of the GWAS data for both traits only identified a few shared loci [2], and LDSC did not detect significant similarity between the two traits (r=-0.002, p=0.945)[5, 39]. We detected a highly significant similarity between the two traits with p-value = 4.19e-10, with Height ranked as the 4th most similar trait to Schizophrenia, even before Bipolar Disorder, which is known to be similar to Schizophrenia (Figure 2A).

We found that there is no significant overlap between the top genes (FDR<0.1) associated with the two traits. This suggests that the strong global similarity we detected is distributed through

genes with weaker associations. Indeed, we found statistically significant overlap when the threshold is lowered: with a cutoff of p= 0.0027 (z>3), 190 genes are associated with Schizophrenia and 278 genes associated with Height, and 17 are in common, with 1.7 expected by chance (p-value = 4.5e-11) (Figure 2B).

Out of the 17 overlapping genes, several are known to be associated with one or both traits. PPARD has been reported to play a protective role in Schizophrenia development[40], and its polymorphism has been associated with reduced Height [41]. SBNO1 is associated with Height and Head Circumference and is important for brain development[42, 43]. MAP3K3 was known to be associated with Height[44]. CNNM2 encodes a metal transporter that plays an important role in magnesium homeostasis and was reported to be involved in the etiology and pathogenesis of Schizophrenia[45]. RP11.724N1.1 is a long noncoding RNA residing in a locus found to associate with Schizophrenia in previous GWAS[46], and its expression was found to be ~3 fold higher in Schizophrenia cohort than control[47]. HIST1H2AC encodes a histone 2A protein and its downregulation is observed in Schizophrenia patients [48]. These 17 genes are located on 6 different chromosomes, providing multiple lines of evidence supporting the genetic similarity between the two traits.

Examination of SNPs that support the shared gene revealed that in general multiple eSNPs align with weak to moderate GWAS peaks, and collectively they give a strong signal. There are many instances where a gene's eQTL profile align to the GWAS profiles of the two traits through different SNPs not in LD, and often located in different chromosomes. For example, for RPL23, one eSNP on chromosome 2 is significantly associated with Schizophrenia but not Height, while another eSNP on chromosome 17 is significantly associated with Height but not Schizophrenia (Figure 2C). Similarly, one of eSNPs for RP11-724N1.1 aligns with Schizophrenia GWAS while another one aligns with Height GWAS; both SNPs are on chromosome 10 but no in LD. These examples illustrate that gene-based approach can identify similarities otherwise invisible to SNP based similarity measures.

To detect groups of functionally related genes that support the similarity between the two traits, we applied PPCA to the two gene association vectors representing Schizophrenia and Height. This analysis identified 9 GO categories with z > 5 with a range of different biological functions, with a general theme on cell differentiation and growth (the top five GO terms are presented in Figure 3).

The top GO category is "skeletal muscle tissue regeneration", with PPARD (peroxisome proliferator activator receptor delta) as a representative gene. PPARD encodes a ligand-activated transcription factor that regulates lipid metabolism. It is a key metabolic regulator in skeletal muscle, and has been implicated in muscle growth due to exercise[49, 50]. Previous studies also suggested that PPARD plays a protective role against the development of Schizophrenia[40]. Interestingly, the functional analysis of SNPs passing genome-wide significance threshold in Schizophrenia GWAS revealed that they were enriched in enhancers specific for neurons and immune cells, as well as skeletal muscles[46]. These combined pieces of evidence suggest that genes involved in the regulation of skeletal muscle growth and differentiation may influence both the growth of the human body as well as the development of Schizophrenia. While the connection to Height may seem intuitive, the mechanism connecting these genes to

Schizophrenia is still obscured. One possibility is that some of these genes may have pleiotropic effect on both muscle and neural development. E.g., PPARD's role on lipid metabolism is important for muscle regeneration[51], but is also important for lipid metabolism in the brain and the myelination of neurons[52]. Consistent with this notion, PPCA of KEGG pathways revealed a related pathway "nuclear receptors in lipid metabolism and toxicity" with PPARD as one representative.

Other top Go categories associated with both Schizophrenia and Height include "large ribosomal subunit rRNA binding", "magnesium ion homeostasis", "positive regulation of cell proliferation in bone marrow", and "dynactin complex", with RPL23, CNNM2, MAP3K3, and ACTR1A as representative genes respectively. Ribosome is the protein synthesis machine and key to cell growth; thus, the connection to height seems intuitive. Previous studies found reduced protein synthesis in Schizophrenia patient derived cells[53], suggesting that decreased protein synthesis may affect brain development in these patients. CNNM2 is associated with the regulation of serum concentration of magnesium, which is known to correlate with height[54]. Previous studies associated a variant of CNNM2 with Schizophrenia risk and decreased gray matter volume in the bilateral inferior frontal gyri[55]. MAP3K3 was associated with Height in a previous GWAS meta-analysis and a follow up study in Brazilian children[44]. MAP3K3 is known to regulates NF-kB and immune response, and dysregulation of immune system plays an important role in Schizophrenia. ACTR1A is a dynactin complex protein. Dynactin is a protein complex that acts as a co-factor for the microtubule motor cytoplasmic dynein-1. It was reported that dynactin interacts with two schizophrenia associated proteins DTNBP1 (dysbindin) and DISC1 (disrupted in Schizophrenia-1) [56-58]. There was no previous report on the connection between Dynactin with Height.

To summarize, PPCA of GO and KEGG identified genes involved in a broad range of functions supporting the genetic similarity between Schizophrenia and Height, with a theme on growth and body/neural development.

Similarity between AD and Cancer and supporting genes/pathways

Epidemiological studies found that risk for cancer and AD is inversely correlated. E.g., in a prospective longitudinal study, it was found that the risk of developing cancer is less among participants with AD vs. nondemented participants and that the risk of developing AD is less for participants with a history of cancer[59]. It has been speculated that such inverse correlation may involve genes regulate cell growth vs. maintenance and cell survival vs. cell death[60]. However, the genetic basis for the correlation and specific genes/pathways that support the correlation have not been analyzed systematically.

We detected a strong similarity between AD and several types of cancer (breast, prostate, pancreatic), with breast cancer scored as the most similar among them (p-value =3.82e-8) (Figure 1).

We applied PPCA to the AD-Breast Cancer pair, and identified 8 GO terms and 4 KEGG pathway with PPC-zscore > 4 (Figure 3). The top 5 GO terms (representative genes) are

“negative regulation of amyloid-beta formation” (APOE, BIN1), “positive regulation of nitric-oxide synthase activity” (APOE, AKT1, HIF1A), “nucleoside-diphosphatase activity” (ENTPD6, NUDT5), “apoptotic nuclear changes” (SHARPIN), and “regulation of neuronal synaptic plasticity” (APOE, SYT4). Consistent with the GO terms, two most significant KEGG pathways were “hypoxia-inducible factor in the cardiovascular system” (P4HB, HIF1A) and “p53 signaling pathway” (SHISA5, CCND3). Several of the representative genes were previously found to be associated with either AD or cancer, or both. E.g., APOE is the most famous gene for AD, and is also associated with cancer[61]. AKT1 has been reported to govern breast cancer progression in vivo[62] and reactive oxygen species-mediated loss of synaptic Akt1 signaling leads to deficient protein translation early in AD[63].

Two functional themes were found by the PPCA. One is hypoxia response, implicated by both GO (“positive regulation of nitric-oxide synthase activity”) and KEGG (hypoxia-inducible factor in the cardiovascular system). Nitric oxide (NO) is a major signaling and effector molecule mediating the body's response to hypoxia[64]. It is known that tumors grow in an environment that lacks oxygen, and the ability to adapt to a hypoxic environment is a key to cancer survival and proliferation. Hypoxia is also linked to neural inflammation, and cerebral hypoxia has been linked to beta-amyloid formation in AD patients. A better hypoxia response may lead to the suppression of AD pathology[65]. Thus, hypoxia response may serve as one mechanism underlying the inverse correlation between cancer and AD. The second functional theme is p53 signaling and apoptosis, also implicated by both GO and KEGG. P53 is a master regulator of cell fate, including cell cycle, senescence, and apoptosis. It is a cancer suppressor that is most frequently mutated in various cancer cells. Genetic variations that strengthen P53 signaling will better suppress cancer. However, in the context of AD, it may also increase the neuronal death that is a major feature of AD pathology. Thus, P53 pathway/apoptosis may be another mechanism supporting the inverse correlation between cancer and AD.

Clustering of phenotypes based on their similarity scores

As described above, having the pairwise similarity scores allows us to identify related traits, much like a blast search for sequence similarity. A different perspective on this data is to cluster phenotypes together based on their similarity score by mapping the similarity score to a distance (see Methods). The result of this global clustering analysis is a tree in which phenotypes with high similarity scores are placed next to each other in a subtree (Figure 4). The advantage of this approach is that the tree will capture the relationships between the multiple phenotypes in the clusters, although significant pairwise similarity may not be readily visible due to the hierarchical nature of the clustering.

We found that similar/related phenotypes were clustered together as expected. These sub-clusters are formed by 1) GWASs of the same phenotype (e.g., osteoporosis, AD, diabetes, and autism); 2) phenotypes that are closely related and were studied in the same GWASs (e.g., LDL cholesterol, HDL cholesterol, total cholesterol, and Triglyceride; diastolic and systolic blood pressure); 3) phenotypes that are known to share genetic etiology in previous studies (e.g., bipolar disorder (BP), schizophrenia (SZ), the combination of the BP and SZ, and five psychiatric disorders including BP and SZ [66], and 4) phenotypes with strong evidence

supporting their shared mechanisms (e.g., Rheumatoid arthritis, Crohn's disease, and Ulcerative Colitis, due to their common inflammatory/autoimmune nature).

The clustering analysis also revealed a number of sub-trees in which phenotypes with no obvious relations are linked together (Figure 4). Here we discuss two examples: 1) a subtree that contains age at menarche, childhood obesity, body mass index; 2) a subtree with 4 phenotypes: fasting glucose, breast cancer, lung cancer, and prostate cancer. Upon literature search, we found that the groupings were supported by previous epidemiologic observations: 1) a previous study found that girls experiencing early menarche are significantly overweight/obese [67]; 2) Patients with high fasting glucose level were reported to have excess risk of all-cause mortality with non-small cell lung cancer[68]. Similarly, high glucose promotes breast cancer proliferation and metastasis by impairing angiotensinogen expression[69]. These were the observed phenotype correlations, and our analyses indicate that they share a common genetic basis.

We next analyzed the shared genes/pathways between the phenotypes in these subtrees.
For the age-at-menarche/body-mass-index/childhood-obesity subtree, we found overlapping genes in each of the three pairwise comparisons. Multiple genes are associated with age-at-menarche and body-mass-index with a theme on cell growth and proliferation. For example, SMAD3 is significantly associated with both age-at-menarche and body-mass-index and is a pivotal intracellular nuclear effector of the TGF-beta (transforming growth factor beta) signaling which plays a critical role in regulating cell grow, differentiation and development[70]. MAP2K5 (Mitogen-Activated Protein Kinase Kinase 5) is also significantly associated with both the phenotypes and the signaling cascade mediated by this kinase is involved in growth factor stimulated cell growth and muscle differentiation [71]. Genes simultaneously associated with both childhood-obesity and body-mass-index include TFAP2B and CENPO. TFAP2B is a transcription factor of AP-2 family that stimulates cell proliferation and suppresses terminal differentiation of specific cell types during embryonic development [72]. CENPO encodes a component of the interphase centromere complex which is required for bipolar spindle assembly, chromosome segregation, and checkpoint signaling during mitosis[73]. Thus, these traits share a set of genes with a common theme on cell growth/division and development.

For the subtree with 4 phenotypes including fasting glucose, breast cancer, prostate cancer and lung cancer, there is no significant overlap among the top associated genes. We performed PPCA to identify shared functional groups. The analysis identified a number of shared GO categories between pairs of traits, some of them common across all 4 phenotypes (Figure S6).

The most frequent GO term that appears in all phenotype pairs is "tail-anchored membrane protein insertion into ER". This GO term includes several genes (such as SGTA and BAG6) involved in the post-translational pathway through which tail-anchored membrane proteins are recognized and targeted to the ER membrane[74]. These genes play important roles in many cellular processes including protein quality control and cell cycle regulation, apoptosis, and DNA damage response. SGTA over-expression was involved in the pathogenesis of breast cancer [75], and BAG6 is involved in the control of apoptosis and is required for the p300-mediated acetylation of P53 upon DNA damage, linking them to cancer risk[76, 77]. BAG6 also participates in endoplasmic-reticulum-associated degradation (ERAD)[78], which is important for protein quality control and glucose-simulated insulin secretion in pancreatic beta cells[79], potentially connecting it to the fasting glucose phenotype.

Among other significant GO terms from the pairwise comparisons, "NAD(P)H oxidase activity" connects all three cancer types (breast cancer, prostate cancer, lung cancer). NAD(P)H oxidases are known to express at a high level and produce ROS in cancer cells. Over-expression of these enzymes (possibly induced by inflammatory signal) has been linked to tissue injury and DNA damage from ROS that accompany pre-malignant conditions, contributing to the initiation and progression of a wide range of solid and hematopoietic malignances[80]. MICAL1 is a major contributing gene within this pathway. It is reported to control cell invasive phenotype via regulating oxidative stress in breast cancer cells[81]. A recent study also found it be regulated by NEDD9 that facilitates hypoxia-induced gastric cancer cell migration in a Rac1-dependent manner[82].

Discussion

We have developed a computational framework to analyze the genetic similarity between complex traits and delineate the potential mechanisms underlying the similarity by integrating GWAS of multiple phenotypes with eQTL data. We demonstrated the approach by a global analysis of published GWASs covering 59 complex human phenotypes. Our analysis revealed expected as well as unexpected relationships between phenotypes and suggested specific mechanisms that connect these phenotypes.

Compared with genetic similarity analysis methods using GWAS data at the SNP level, our gene-based similarity measure has several advantages. First, comparing the genetic profiles at the gene level can yield a stronger signal, as multiple SNPs can converge on the same gene. Even if the GWAS peaks of two different phenotypes do not align by proximity, they can be "aligned" through a gene, i.e., their effects on the phenotype can be mediated through the same gene. Thus, our method can be more sensitive in detecting genetic similarity between traits. Our method identified many trait pairs with a strong similarity that were not detected by SNP based similarity measures such as LDSC regression[5] (Figure 1). Examples include height vs. schizophrenia and AD vs. coronary artery diseases, both were observed in previous epidemiological studies[1, 83], and Crohn's disease vs. Rheumatoid Arthritis, which are known to share inflammation/immune origin. Secondly, genetic similarity analysis at the gene level readily suggests potential mechanisms connecting different phenotypes. We used two different approaches to delineate genes/pathways co-associated with both traits; one is to analyze the overlapping genes passing a certain p-value cutoff for each of the two traits, another one is to identify a group of functionally related genes using PPCA. These approaches lead to the identification of specific genes/pathways shared by the two traits and provide potential mechanistic connections.

We note that our method may also miss some relations detected by LDSC. This can happen, e.g., if the two phenotypes share some GWAS SNPs that do not align to eSNPs, either because the SNPs do not influence the phenotype by influencing mRNA expression, or the eQTL data (which is often noisy and incomplete) failed to capture the eSNPs. In a systematic comparison, we found that among the trait pairs analyzed by both LDSC and our method, with a p-value cutoff of 1.0e-

3, our method identified 146 pairs while LDSC identified 93, with 60 pairs overlap (p-value for the overlap = 4.9e-9) (Dataset S4). This suggests that although gene-based method has higher sensitivity, the two methods are complementary.

One limitation of the gene-based method is that it does not predict the direction by which the two traits are correlated, while LDSC does predict a direction. Predicting direction by gene-based approach is complicated by several factors: 1) the p-value of association between a gene and a phenotype is typically supported by multiple SNPs, and when considering the relative sign between the effect on gene expression and the effect on phenotype, these SNPs can have either concordant or discordant directions. This is a previously observed phenomenon with an unknown mechanism[12] and makes it difficult to assign a direction for the gene-phenotype association, although it can be analyzed case by case. 2) the directional information for SNPs are not always available. We do not have directional information for most of the eQTL data we used except for GTEx. GTEx is systematic, but due to sample size limitation does not give reliable trans-eQTL information. The simplicity of the gene-based method is that it only needs p-values of association for the eQTL and GWAS data. Future work is needed to factor in directional information as more systematic and larger scale eQTL data become available.

A key step in our gene-based method is to translate the SNP-phenotype association profile to the gene-phenotype association profile. For this purpose, we developed the Sherlock-II algorithm, which is a second generation of the Sherlock algorithm we developed before. Sherlock-II kept the basic spirit of Sherlock in that it integrates information from all the SNPs (SNPs with both strong and weak GWAS p-values, and both cis and trans to a gene), but is more robust and efficient than Sherlock. The integration over all SNPs is essential for translating the global SNP-phenotype profile to gene-phenotype profile while preserving as much information as possible. This is important for detecting signals in the gene-phenotype association and for the success of the gene-based similarity measure.

A number of elegant GWAS vs. eQTL alignment tools have been developed, most of them focus on the colocalization between strong GWAS peaks and cis-eQTL peaks. For example, eCAVIER uses a Bayesian model to colocalize eQTL and GWAS peaks by considering local LD structure[11]. SMR computes the effect size from gene to phenotype and tests for consistency across strong cis-eSNPs that influence the same gene[12]. Both methods start with a strong GWAS/eQTL peak passing genome-wide significance threshold and analyze the region around it, leading to high specificity of the predictions. However, they infer genes based on mostly a single strong signal instead of integrating signals from multiple SNPs. Compared with these two methods, Sherlock-II detects gene-phenotype association with higher sensitivity, especially when the signals in the original GWAS are weak. For example, from the autism GWAS data, Sherlock-II identified 10 genes with FDR<0.1 (supplementary data1), while eCAVIAR and SMR could not identify any gene because none of the autism GWAS signals passed the genome-wide significance threshold[26]. We found literature support for several of the 10 genes, including ENPP4, LEPR, and MTHFR [26, 27, 29, 84].

It is important to note that the data we obtained are co-associations between genes and phenotypes (supported by multiple GWAS and eQTL SNPs), not necessarily causal relations. When a gene G is found to be co-associated with phenotypes A and B, there are several different

possible scenarios, including 1) G influences both A and B, and 2) G influence one of the phenotypes which in turn influences another phenotype. In some of the cases we discussed, such as the co-association of genes involved in apoptosis with cancer and AD, it is likely that 1) is the correct model, while in some other cases it might be difficult to distinguish the models. For example, genes involved in ER membrane insertion may influence insulin secretion and fasting glucose level, which in turn affects cancer risk. Clearly more experiments/analyses are required to sort out the causality for the genes/pathways we identified to be co-associated with many trait pairs.

Methods

Measuring genetic similarity between two traits using their gene-phenotype association profiles

Vector representation of the gene-phenotype association profile

We first translate the SNP-phenotype association profile (p-values of association for all the SNPs with the phenotype) into a gene-phenotype association profile (p-values of association for all the genes with the phenotype) using the Sherlock-II algorithm (see the last section in Methods). To measure genetic similarity between two phenotypes, we represent the association profile of a phenotype *t* by a vector of dimension *N* (the total number of genes in the genome), $V(t) = [Z_1(t), Z_2(t), \dots, Z_N(t)]$ where $Z_i(t)$ can be calculated from the p-value of association between gene *i* and trait *t*, by inverting the following equation: $P_i(t) = Prob(|X| \geq Z_i(t))$, where *X* is a random variable with standard normal distribution. This p-value to z transformation is necessary to emphasize the contributions of genes with small p-values (indicating significant association between the expression level of the gene and the phenotype) when measuring the distance between a pair of gene-phenotype association vectors. The z score is unsigned, since we have p-value of association for the gene calculated from Sherlock-II without direction (see Discussion). We experimented with other types of vector representation, e.g., $-log_{10}(P_i(t))$, and find very similar results.

Gene-based genetic similarity score between two phenotypes

Using the vector representation for the traits, we define a genetic similarity score between a pair of traits t1 and t2 as $S(t_1, t_2) = \left[\overline{E_{1,2^*}} - E\big(V(t_1), V(t_2)\big)\right] / \sigma_{E_{1,2^*}}$ Where $E\big(V(t_1), V(t_2)\big)$ is the Euclidean distance between the two gene association vectors $V(t_1)$ and $V(t_2)$, $\overline{E_{1,2^*}}$ and $\sigma_{E_{1,2^*}}$ are the mean and standard deviation of a background distribution of Euclidean distances assuming that the two vectors are un-correlated, generated by randomly permuting the components of the second vector (denoted as 2^*). A pair of phenotypes similar to each other are expected to have similar association vectors, and will give rise to smaller distance and larger similarity score. The normalization by the background distribution (measured as the number of standard deviations away from the mean) is essential as it takes out the contribution to the distance due to the features of each individual vectors, since we are interested in the similarity

between the two vectors, not the vectors themselves. We experimented with different distance metric (such as Manhattan distance) and found very similar results (Figure S4).

Assessing the statistical significance of the similarity score

To assess the statistical significance of the similarity score, we generated a null distribution of similarity scores by comparing a set of "fake" GWAS data with the real GWAS data we have. Fake GWAS data were generated by randomly select "cases" and "controls" from the subjects from the 1000 genome projects. We expect that these fake GWASs should not be similar to any of the real GWASs we analyzed, and thus the similarity scores obtained by comparing the fake with the real GWASs form a background distribution from which we can calculate the statistical significance of the observed similarity score.

We generated 500 fake GWASs (randomly selecting 500 cases and 500 controls) and used Plink[85] to compute the summary statistics. We then applied Sherlock-II to the summary statistics and generated gene association vectors for these fake GWASs using the same eQTL data we used for analyzing the real GWASs. We calculated the pairwise similarity score between a real GWAS and a fake GWAS using the same equation above, and generated 69 × 500 similarity scores $S(t_i, r_j) = \left[\overline{E_{i,j^*}} - E\left(V(t_i), V(r_j)\right)\right] / \sigma_{E_{i,j^*}}$ between 69 real GWASs and 500 fake GWASs, where r_j is the j-th fake GWAS. The distribution of these scores is well fitted to a normal distribution, with a mean $\overline{S}$ and standard deviation σ_S. For an observed similarity score S, we calculate a z score $Z(t_1, t_2) = [S(t_1, t_2) - \overline{S}]/\sigma_s$, and convert it to a p-value based a one tailed test using standard normal distribution.

Partial-Pearson-Correlation Analysis (PCCA) to identify sets of functionally related genes that contribute to the genetic similarity

We employed a simple statistical approach to detect gene ontology terms that contribute significantly to the genetic similarity between two different phenotypes. Gene Ontology gene annotations were downloaded from Ensembl. We removed the GO terms with less than 5 genes (not enough statistics) and more than 100 genes (too non-specific) from the analysis, after which 6796 GO terms were used. For a specific pair of phenotypes, we calculate a Partial-Pearson-Correlation-Analysis zscore (*Z*) for each of the 6796 GO terms using the following formula (previously developed for analyzing similarity between two gene expression profiles [8])

$$Z_J = \frac{\sum_{i \in J}(x_i - \mu_x)\left(y_i - \mu_y\right)}{\sigma_x \sigma_y \sqrt{n_J}}$$

where x_i and y_i are the z values associated with the two phenotypes for gene *i* (calculated from the SherlockII p-values for gene-phenotype association), μ_x, σ_x and μ_y, σ_y are the mean and standard deviation for x and y respectively (calculated from all genes), J is the gene set defined by the GO term, and n_J is the number of genes in the GO term. This formula is similar to the one for calculating Pearson correlation (with a normalization), except that the summation is partial – only for the genes in the subset. We argued that the null distribution for should be standard normal distribution, and verified that this is indeed the case by randomly sample all phenotypes

pairs and generate for all the GO terms (Figure S7).

Hierarchical clustering of phenotypes

To performed hierarchical clustering of traits based on their pairwise similarity scores, we first map the similarity score to a distance between 0 and 1 using the following equation:

$d(t_1, t_2) = 1/[1 + \max(0,\ S(t_1, t_2) - S_M)]$, where S_M is the median of all pairwise similarity scores. This means that any pair with similarity score smaller than the median will get a maximum distance of 1. For similarity score much larger than the median, the distance approaches 0. We then apply hierarchical clustering algorithm to this distance matrix, using average distance to measure the distance between composite nodes. We verified that the structure of the tree is not sensitive to the way the distance between composite nodes is measured. E.g, using weighted distance yielded a very similar tree (Figure S5)

Sherlock-II algorithm: translating SNP-phenotype association profile to gene-phenotype association profile

Sherlock-II detects the association between an intermediate molecular trait (IMT) and a phenotype by comparing the GWAS SNPs with the QTL SNPs of the IMT (eQTL or metabolite-QTL), with the assumption that if an IMT is causal to the phenotype, SNPs that influence the IMT should also influence the phenotype, thus the QTL profile of the IMT should have significant overlap with the GWAS profile of the phenotype. To test the significance of the overlap, Sherlock-II first selects the QTL SNPs of an IMT that pass a threshold, aligns these SNPs to the corresponding SNPs in the GWAS, and test whether the aligned GWAS SNPs have p-values significantly better than those picked at random.

When testing the significance of a selected set of loci in a given GWAS, Sherlock-II uses a simple scheme to compute a score for the n p-values of the GWAS SNPs that best tag these loci. It then uses a convolution-based approach to construct an empirical null distribution for scores involving n p-values drawn randomly from the full set of GWAS p-values. In certain instances, particularly with eQTL for many genes, this test will be performed repeatedly for all intermediate molecular traits in a study. If dependence between the sets of IMTs is detected — for example, due to pleiotropic loci that regulate many genes — a provision of adjusting the null distribution is included. Through extensive testing and simulation, we show that our empirical strategy produces p-values that are highly resistant to inflation. In this section, we describe the individual components of the Sherlock-II method in detail in the order in which they occur.

Identifying Tag SNPs for Independent Blocks

Sherlock-II statistics is based on the assumption that the collection of SNPs tested are tagging independently segregating haplotype blocks in both the GWAS and eQTL cohorts. Understanding the true block structure of the cohorts is critical to accurately estimating gene significance. The inadvertent inclusion of dependent blocks can inflate the significance of genes by over-counting true associations; conversely, a conservative approach that enforces wide separation between tagging SNPs may miss many independent loci. The preferred approach to identifying independent blocks requires genotypes for both the GWAS and functional data to

define block boundaries based on regions of rapid decay in linkage-disequilibrium (LD). However, since the difficulty of obtaining genotypes across a large number of GWAS studies makes this impractical, we constructed a database of LD using Caucasian cohorts from the 1000 Genomes Project. We use PLINK (v1.07) to compute r^2 linkage between common SNPs for 379 individuals in the CEPH, CEU, TSI, GBR, FIN, and IBS cohorts of 1000 Genomes release v3.20101123. This permits the alignment of associated loci in the functional data to corresponding SNPs in GWAS data, and it enables the identification of independently-segregating tag SNPs in the combined data for use in the statistical test.

We first match the GWAS SNPs to all the loci present in a given functional data set that pass a specific threshold. For the eQTL data used here, the nominal threshold is typically eSNP p-values < 10^{-5}. If matching SNPs are not present in both data sets, we use the closest GWAS SNPs in r^2 LD > 0.85, if available. We then use an agglomerative clustering approach to identify non-overlapping blocks of SNPs in r^2 LD < 0.2. From each block, the functional SNP with the strongest association is selected as a tag, and its corresponding GWAS SNP is included in the set subjected to statistical test. Since discrepancies between the various cohorts — GWAS, functional, and linkage — are inevitable, a minimum 100 kb distance between tag SNPs is enforced. In addition, we exclude SNPs from the human leukocyte antigen (HLA) region between 6p22.1 to 6p21.3 due to its complex linkage structure.

Core Statistical Method

Once a set of SNPs related to a specific molecular trait is identified, we compute a score s by simply combining their GWAS log p-values:

$$s = -\sum_{i=1}^{n} log_{10}(p_i)$$

This converts our collection of SNPs into a scalar quantity that, when referenced against the appropriate null distribution, indicates the statistical significance of selecting this subset from the pool of all independent GWAS loci. Our approach is analogous to Fisher's combined probability test using the scoring function $-2\sum_{i=1}^{n} log_e(p_i)$ which follows a χ^2 distribution with 2n degrees of freedom when p-values are drawn from a uniform distribution on (0,1). With our method, instead of assuming a particular form for the distribution of GWAS p-values, we use discrete convolution to compute an empirically-derived distribution of scores when combining n p-values. Since s represents a specific score, we let S represent a random variable of all possible scores. We form the probability mass function (PMF) of S using bins of width b (log scale), where the probability of scores in the range $0 \leq S < b$ is the first element, $b \leq S < 2b$ is the second element, and so forth. Thus, $f_n[k] = P(kb \leq S < (k+1)b)$ is k-th element of the PMF for a score S comprising n independent GWAS p-values. For the simplest case of scores involving only a single SNP, the PMF f_1 is essentially a normalized histogram of p-values for all independent (unlinked) SNPs in the GWAS that aligned to an eQTL SNP across all genes:

$$f_1[k] = \frac{1}{N}\sum_{i=1}^{N} I[kb \leq -log_{10}(p_i) < (k+1)b]$$

where I is the indicator function, N is the total number of SNPs, b is the bin width, and k is the array index within the histogram. For computational efficiency, the minimum GWAS p-value is typically truncated at 10^{-9}, well below genome-wide significance, yielding 900 bins when spacing b = 0.01 is used. This single-locus PMF forms the basis from which null distributions for any arbitrary number of functionally-related loci are formed. Since the sum of two independent random variables has a PDF equal to the convolution of their individual probability distributions, scores involving two GWAS p-values have PMF $f_2 = f_1 \circledast f_1$. In the general case for any value of n > 1, the PMF f_n is formed from n-1 recursive convolutions of f_1:

$$f_n[k] = \sum_{l=0}^{k} f_1[l] \cdot f_{n\text{-}1}[k\text{-}l]$$

where again k is the array index of probability mass function. For the special case of a truly uniform p-value distribution, we have an exact continuous solution to the recursive convolution for n p-values (which can be obtained from the chi-square distribution).

To motivate the use of our empirical approach, we compare this theoretical result to our empirical method for GWAS of differing power. For low-powered association studies with p-value distributions close to uniform, the tail of the empirical distribution is similar to the tail of the theoretical distribution, leading to roughly identical estimates of significance for a given score (see below). For well-powered meta-analyses that appear to contain inflation from various sources, significant differences in the distribution tails can lead to an appreciable overestimation of significance for high scores when using the theoretical distribution. With the empirical distribution derived from the real GWAS data, Sherlock-II is insensitive to the input GWAS inflation.

Computing the significance of a gene with score s

We used p-value to characterize the significance of the gene score s by summing the tail of the background distribution f_n , where n is the number of independent GWAS loci supporting this gene. The tail is defined as the bins correspond to S that is great or equal to observe gene score s.

$$P(S \geq s) = \sum_{k=k_s}^{\infty} f_n[k], \text{ while } k_s b \leq s < (k_s + 1)b$$

Correcting for Pleiotropic and Sampling Effect

Another source of inflation stems from a lack of true independence between molecular traits across a given functional dataset. For example, pleiotropic loci may appear to regulate hundreds of genes in the eQTL data for a given tissue. In our simulations, these loci may inflate the output test statistic due to chance alignment with significant GWAS SNPs; in our tests, this occurs at an appreciable level in perhaps five percent of the cases. Since our method permits the use of a unique distribution f_1 for each p-value added to the score, it enables a simple scheme for reducing inflation by adjusting each distribution based on the number of molecular traits affected by the locus. For the eQTL example, this involves conditioning the GWAS distribution based on the number of genes influenced by a locus across the entire eQTL data set. When chance alignment of pleiotropic loci and significant GWAS SNPs occurs, the overrepresentation of small p-values is reflected in the null distribution for affected genes. Thus, in practice, the actual

distribution incorporated into a Sherlock-II score with eQTL data is conditioned on the number of genes that are regulated by the same locus:

$$f_{1|c\in C_{family}}[k] = \frac{\sum_{i=1}^{N} I[kb \leq \text{-}log_{10}(p_i) < (k+1)b] \cap I[C_i \in C_{family}]}{\sum_{i=1}^{n} I[C_i \in C_{family}]}$$

where I is the indicator function, N is the total number of SNPs, k is the index, b is the bin width, and the ith SNP is assigned gene count C_i while the locus in question has count c. C_{family} is several non-overlapping sets, each including a range of gene count c. For example, $C_{family6}$ includes all SNPs that regulate 6 to 10 genes. Now the distribution contains only the p-values of SNPs that belong to the same count family. With this change, instead of computing the significance of a set of SNPs drawn from GWAS, our method computes the significance of a set of SNPs drawn given their individual pleiotropies count family. In our simulations, this results in a minor change in the significance and rank order of output genes in most cases but prevents inflation whenever chance alignment of significant GWAS SNPs and pleiotropic loci occurs (Figure S8). Specifically, we penalize SNPs regulating the expression of a large number of genes.

References

1. Zammit, S., et al., *Height and body mass index in young adulthood and risk of schizophrenia: a longitudinal study of 1 347 520 Swedish men.* Acta Psychiatr Scand, 2007. **116**(5): p. 378-85.
2. Bacanu, S.A., X. Chen, and K.S. Kendler, *The genetic overlap between schizophrenia and height.* Schizophr Res, 2013. **151**(1-3): p. 226-8.
3. Pickrell, J.K., et al., *Detection and interpretation of shared genetic influences on 42 human traits.* Nat Genet, 2016. **48**(7): p. 709-17.
4. Turley, P., et al., *Multi-trait analysis of genome-wide association summary statistics using MTAG.* Nat Genet, 2018. **50**(2): p. 229-237.
5. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-41.
6. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
7. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
8. McCarroll, S.A., et al., *Comparing genomic expression patterns across species identifies shared transcriptional profile in aging.* Nat Genet, 2004. **36**(2): p. 197-204.
9. *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
10. Giambartolomei, C., et al., *Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.* PLoS Genet, 2014. **10**(5): p. e1004383.
11. Hormozdiari, F., et al., *Colocalization of GWAS and eQTL Signals Detects Target Genes.* Am J Hum Genet, 2016. **99**(6): p. 1245-1260.
12. Zhu, Z., et al., *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet, 2016. **48**(5): p. 481-7.

13. He, X., et al., *Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS.* Am J Hum Genet, 2013. **92**(5): p. 667-80.
14. Wen, X., R. Pique-Regi, and F. Luca, *Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization.* PLoS Genet, 2017. **13**(3): p. e1006646.
15. Barbeira, A.N., et al., *Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics.* Nat Commun, 2018. **9**(1): p. 1825.
16. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nat Genet, 2015. **47**(9): p. 1091-8.
17. Dixon, A.L., et al., *A genome-wide association study of global gene expression.* Nat Genet, 2007. **39**(10): p. 1202-7.
18. Duan, S., et al., *Genetic architecture of transcript-level variation in humans.* Am J Hum Genet, 2008. **82**(5): p. 1101-13.
19. Liang, L., et al., *A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines.* Genome Res, 2013. **23**(4): p. 716-26.
20. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nat Genet, 2012. **44**(10): p. 1084-9.
21. Myers, A.J., et al., *A survey of genetic human cortical gene expression.* Nat Genet, 2007. **39**(12): p. 1494-9.
22. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.
23. Wright, F.A., et al., *Heritability and genomics of gene expression in peripheral blood.* Nat Genet, 2014. **46**(5): p. 430-7.
24. Zeller, T., et al., *Genetics and beyond--the transcriptome of human monocytes and disease susceptibility.* PLoS One, 2010. **5**(5): p. e10693.
25. *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
26. *Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia.* Mol Autism, 2017. **8**: p. 21.
27. Zhang, L., et al., *Genome-wide association study and identification of chromosomal enhancer maps in multiple brain regions related to autism spectrum disorder.* Autism Res, 2019. **12**(1): p. 26-32.
28. Velmeshev, D., et al., *Cell-Type-Specific Analysis of Molecular Pathology in Autism Identifies Common Genes and Pathways Affected Across Neocortical Regions.* Mol Neurobiol, 2020. **57**(5): p. 2279-2289.
29. You, Y.H., et al., *MicroRNA-153 promotes brain-derived neurotrophic factor and hippocampal neuron proliferation to alleviate autism symptoms through inhibition of JAK-STAT pathway by LEPR.* Biosci Rep, 2019. **39**(6).
30. Rai, V., *Association of methylenetetrahydrofolate reductase (MTHFR) gene C677T polymorphism with autism: evidence of genetic susceptibility.* Metab Brain Dis, 2016. **31**(4): p. 727-35.
31. Long, T., et al., *Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites.* Nat Genet, 2017. **49**(4): p. 568-578.
32. Korb, A., H. Pavenstädt, and T. Pap, *Cell death in rheumatoid arthritis.* Apoptosis, 2009. **14**(4): p. 447-54.

33. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.* Nat Genet, 2010. **42**(12): p. 1118-25.
34. Wahlbeck, K., et al., *Association of schizophrenia with low maternal body mass index, small size at birth, and thinness during childhood.* Arch Gen Psychiatry, 2001. **58**(1): p. 48-52.
35. Weiser, M., et al., *Body mass index and future schizophrenia in Israeli male adolescents.* J Clin Psychiatry, 2004. **65**(11): p. 1546-9.
36. Gunnell, D., et al., *The association of fetal and childhood growth with risk of schizophrenia. Cohort study of 720,000 Swedish men and women.* Schizophr Res, 2005. **79**(2-3): p. 315-22.
37. Sørensen, H.J., et al., *Height, weight and body mass index in early adulthood and risk of schizophrenia.* Acta Psychiatr Scand, 2006. **114**(1): p. 49-54.
38. Wyatt, R.J., et al., *Height, weight and body mass index (BMI) in psychiatrically ill US Armed Forces personnel.* Psychol Med, 2003. **33**(2): p. 363-8.
39. Duncan, L.E., et al., *Genetic Correlation Profile of Schizophrenia Mirrors Epidemiological Results and Suggests Link Between Polygenic and Rare Variant (22q11.2) Cases of Schizophrenia.* Schizophr Bull, 2018. **44**(6): p. 1350-1361.
40. Li, X., et al., *PPARD May Play a Protective Role against the Development of Schizophrenia.* PPAR Res, 2020. **2020**: p. 3480412.
41. Burch, L.R., et al., *A single nucleotide polymorphism on exon-4 of the gene encoding PPARdelta is associated with reduced height in adults and children.* J Clin Endocrinol Metab, 2009. **94**(7): p. 2587-93.
42. Takano, A., et al., *Function of strawberry notch family genes in the zebrafish brain development.* Kobe J Med Sci, 2011. **56**(5): p. E220-30.
43. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
44. Fontenele, E.G., et al., *Association Study of GWAS-Derived Loci with Height in Brazilian Children: Importance of MAP3K3, MMP24 and IGF1R Polymorphisms for Height Variation.* Horm Res Paediatr, 2015. **84**(4): p. 248-53.
45. Guan, F., et al., *Two-stage replication of previous genome-wide association studies of AS3MT-CNNM2-NT5C2 gene cluster region in a large schizophrenia case-control sample from Han Chinese population.* Schizophr Res, 2016. **176**(2-3): p. 125-130.
46. *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-7.
47. Liu, Y., et al., *Non-coding RNA dysregulation in the amygdala region of schizophrenia patients contributes to the pathogenesis of the disease.* Transl Psychiatry, 2018. **8**(1): p. 44.
48. Fries, G.R., et al., *Genome-wide expression in veterans with schizophrenia further validates the immune hypothesis for schizophrenia.* Schizophr Res, 2018. **192**: p. 255-261.
49. Zizola, C., et al., *Activation of PPARδ signaling improves skeletal muscle oxidative metabolism and endurance function in an animal model of ischemic left ventricular dysfunction.* Am J Physiol Heart Circ Physiol, 2015. **308**(9): p. H1078-85.
50. Ehrenborg, E. and A. Krook, *Regulation of skeletal muscle physiology and metabolism by peroxisome proliferator-activated receptor delta.* Pharmacol Rev, 2009. **61**(3): p. 373-93.

51. Angione, A.R., et al., *PPARδ regulates satellite cell proliferation and skeletal muscle regeneration.* Skelet Muscle, 2011. **1**(1): p. 33.
52. Dickey, A.S., et al., *PPAR-δ is repressed in Huntington's disease, is required for normal neuronal function and can be targeted therapeutically.* Nat Med, 2016. **22**(1): p. 37-45.
53. English, J.A., et al., *Reduced protein synthesis in schizophrenia patient-derived olfactory cells.* Transl Psychiatry, 2015. **5**(10): p. e663.
54. Yin, Y., et al., *Evaluation of the Relationship Between Height and Zinc, Copper, Iron, Calcium, and Magnesium Levels in Healthy Young Children in Beijing, China.* Biol Trace Elem Res, 2017. **176**(2): p. 244-250.
55. Ohi, K., et al., *The impact of the genome-wide supported variant in the cyclin M2 gene on gray matter morphology in schizophrenia.* Behav Brain Funct, 2013. **9**: p. 40.
56. Wang, H., et al., *Dysbindin-1 Involvement in the Etiology of Schizophrenia.* Int J Mol Sci, 2017. **18**(10).
57. Dahoun, T., et al., *The impact of Disrupted-in-Schizophrenia 1 (DISC1) on the dopaminergic system: a systematic review.* Transl Psychiatry, 2017. **7**(1): p. e1015.
58. Nakata, K., et al., *DISC1 splice variants are upregulated in schizophrenia and associated with risk polymorphisms.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15873-8.
59. Roe, C.M., et al., *Alzheimer disease and cancer.* Neurology, 2005. **64**(5): p. 895-8.
60. Behrens, M.I., C. Lendon, and C.M. Roe, *A common biological mechanism in cancer and Alzheimer's disease?* Curr Alzheimer Res, 2009. **6**(3): p. 196-204.
61. Sakashita, K., et al., *Clinical significance of ApoE expression in human gastric cancer.* Oncol Rep, 2008. **20**(6): p. 1313-9.
62. Ju, X., et al., *Akt1 governs breast cancer progression in vivo.* Proc Natl Acad Sci U S A, 2007. **104**(18): p. 7438-43.
63. Ahmad, F., et al., *Reactive Oxygen Species-Mediated Loss of Synaptic Akt1 Signaling Leads to Deficient Activity-Dependent Protein Translation Early in Alzheimer's Disease.* Antioxid Redox Signal, 2017. **27**(16): p. 1269-1280.
64. Umbrello, M., et al., *The key role of nitric oxide in hypoxia: hypoxic vasodilation and energy supply-demand matching.* Antioxid Redox Signal, 2013. **19**(14): p. 1690-710.
65. Lall, R., R. Mohammed, and U. Ojha, *What are the links between hypoxia and Alzheimer's disease?* Neuropsychiatr Dis Treat, 2019. **15**: p. 1343-1354.
66. Lee, S.H., et al., *Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.* Nat Genet, 2013. **45**(9): p. 984-94.
67. Bralic, I., et al., *Association of early menarche age and overweight/obesity.* J Pediatr Endocrinol Metab, 2012. **25**(1-2): p. 57-62.
68. Luo, J., Y.J. Chen, and L.J. Chang, *Fasting blood glucose level and prognosis in non-small cell lung cancer (NSCLC) patients.* Lung Cancer, 2012. **76**(2): p. 242-7.
69. Sun, S., et al., *High glucose promotes breast cancer proliferation and metastasis by impairing angiotensinogen expression.* Biosci Rep, 2019. **39**(6).
70. Kingsley, D.M., *The TGF-beta superfamily: new members, new receptors, and new genetic tests of function in different organisms.* Genes Dev, 1994. **8**(2): p. 133-46.
71. Katz, M., I. Amit, and Y. Yarden, *Regulation of MAPKs by growth factors and receptor tyrosine kinases.* Biochim Biophys Acta, 2007. **1773**(8): p. 1161-76.
72. Eckert, D., et al., *The AP-2 family of transcription factors.* Genome Biol, 2005. **6**(13): p. 246.

73. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
74. Leznicki, P., et al., *The association of BAG6 with SGTA and tail-anchored proteins.* PLoS One, 2013. **8**(3): p. e59590.
75. Zhu, T., et al., *Expression and prognostic role of SGTA in human breast carcinoma correlates with tumor cell proliferation.* J Mol Histol, 2014. **45**(6): p. 665-77.
76. Sasaki, T., et al., *HLA-B-associated transcript 3 (Bat3)/Scythe is essential for p300-mediated acetylation of p53.* Genes Dev, 2007. **21**(7): p. 848-61.
77. Schuldner, M., et al., *Exosome-dependent immune surveillance at the metastatic niche requires BAG6 and CBP/p300-dependent acetylation of p53.* Theranostics, 2019. **9**(21): p. 6047-6062.
78. Payapilly, A. and S. High, *BAG6 regulates the quality control of a polytopic ERAD substrate.* J Cell Sci, 2014. **127**(Pt 13): p. 2898-909.
79. Hu, Y., et al., *Endoplasmic Reticulum-Associated Degradation (ERAD) Has a Critical Role in Supporting Glucose-Stimulated Insulin Secretion in Pancreatic beta-Cells.* Diabetes, 2019. **68**(4): p. 733-746.
80. Roy, K., et al., *NADPH oxidases and cancer.* Clin Sci (Lond), 2015. **128**(12): p. 863-75.
81. Deng, W., et al., *MICAL1 controls cell invasive phenotype via regulating oxidative stress in breast cancer cells.* BMC Cancer, 2016. **16**: p. 489.
82. Zhao, S., et al., *NEDD9 Facilitates Hypoxia-Induced Gastric Cancer Cell Migration via MICAL1 Related Rac1 Activation.* Front Pharmacol, 2019. **10**: p. 291.
83. Sparks, D.L., et al., *Cortical senile plaques in coronary artery disease, aging and Alzheimer's disease.* Neurobiol Aging, 1990. **11**(6): p. 601-7.
84. Pu, D., Y. Shen, and J. Wu, *Association between MTHFR gene polymorphisms and the risk of autism spectrum disorders: a meta-analysis.* Autism Res, 2013. **6**(5): p. 384-92.
85. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

Figure legends

Figure 1. Genetic similarity between pairs of traits detected by the gene-based similarity measure. Shown are top hits to 16 selected traits (first column, bold face). Each row listed the top 8 most similar traits that pass 1.0e-3 p-value cutoff, ranked by their p-values as indicated by the color scale. Pairs covered by a previous analysis using LDSC are labeled with either a check mark, indicating that LDSC also detected the similarity with $p < 0.002$, or with a star, indicating that LDSC did not find significant similarity. Entries colored red indicate unexpected relationships. The numbers associated with the trait name link the trait to a specific GWAS data when there are multiple GWASs for the same trait (see Dataset S3 for the link).

Figure 2. Genetic similarity between Schizophrenia and Height and the supporting genes. Gene-based method detected a strong similarity between Schizophrenia and Height, with a p-value =4.19e-10. A) Height is ranked as the 4th most similar trait to Schizophrenia, before Bipolar disorder. Shown are the similarity scores for Schizophrenia-Height and Schizophrenia-Bipolar disorder (red bars) pairs, and the background distribution of the similarity score generated by comparing fake GWASs with real GWAS data (blue curve). B) 17 genes co-associated with both Schizophrenia and Height, with z>3 for both phenotypes. C) the alignment between the eQTL profile of RPL23 and the GWAS profiles of Schizophrenia and Height. RPL23 is co-associated with both phenotypes, supported by eSNPs that aligned to non-overlapping GWAS peaks (highlighted with red circles) in two different phenotypes.

Figure 3. Partial Pearson Correlation Analysis (PPCA) revealed GO terms and KEGG pathways supporting the genetic similarity between traits. Shown are GO terms with PPCA-zscore >5 and KEGG pathways with PPCA-zscore >4, up to top 5, for Crohn's Disease vs. Rheumatoid Arthritis, AD vs. Breast Cancer, and Height vs. Schizophrenia pairs. Length of the bar and the numbers in the bar indicate PPCA-zscore.

Figure 4. Hierarchical clustering of phenotypes based on their pairwise similarity score. Highlighted branches discussed in the text: 1) GWASs of the same phenotypes (green); 2) phenotypes that are closely related and analyzed in the same GWAS (orange); 3) phenotypes that are known to share genetic etiology (blue); and 4) phenotypes with no obvious relations but co-occurrences were observed based on previous epidemiological studies (red).

Figure S1. Genes significantly associated with Autism detected by Sherlock-II. A) left panel: QQ plot for the original GWAS p-value distribution. Right panel: QQ plot for the gene association p-value distribution. B) the list of 10 genes with FDR<0.1.

Figure S2. Four examples of metabolite-phenotype associations identified by applying Sherlock-II to GWAS data and metabolite-QTL data. QQ plot for metabolite p-value distribution (left panel), QQ plot of original GWAS study (inserts in the left panel), and alignment between the

metabolite-QTL and GWAS profiles (right panel) are presented. (A) Methionine sulfone is associated with Chronic Kidney Disease. (B) Palmitoyl-dihydrosphingomyelin is associated with Alzheimer's disease. (C) Homocitrulline is associated with Autism (data from dbGaP). (D) Gamma-glutamylglutamate is associated with Schizophrenia.

Figure S3. Hierarchical clustering of phenotypes based on their pairwise similarity score, using -log10(P) as the vector representation.

Figure S4. Hierarchical clustering of phenotypes based on their pairwise similarity score, using Manhattan distance as the distance metric.

Figure S5. Hierarchical clustering of phenotypes based on their pairwise similarity score, using weighted distance between composite cluster nodes.

Figure S6. Summary of the top GO categories contributing to the overall similarity between phenotypes for the subtree including four phenotypes (fasting glucose, breast cancer, lung cancer, prostate cancer). Listed are top 5 non-redundant GO categories with highest PPCA-zscore for each pair of the phenotypes. Recurrent GO terms across different pairwise comparisons were highlight with different colors.

Figure S7. Distribution of PPCA-zscores for GO analysis.

Figure S8. SherlockII pleiotropic correction removed inflation from random GWAS.

Figure 1

Height	BMI √	Pubertal Growth Late Adolescence	Age at Menarche √	**Osteoporosis_1**	HDL *	**Schizophrenia_1** *	Cholesterol √	Waist-Hip Ratio adjusted for BMI √
Schizophrenia_1	Bipolar Disorder and Schizophrenia	Five Psychiatric Diseases	**BMI** √	**Height** *	Bipolar Disorder √	Neuroticism_2	**Education Attainment - College** √	**Diastolic Blood Pressure**
Bipolar Disorder	Bipolar Disorder and Schizophrenia	Five Psychiatric Diseases	Schizophrenia_1 √	**AD_1** *	Height *	ADHD *	HDL *	Triglycerides *
Autism_2	Autism_1	**Five Psychiatric Diseases**	Neuroticism_2	BMI *				
Ulcerative Colitis	Crohn's Disease √	Rheumatoid Arthritis *						
Rheumatoid Arthritis	Crohn's Disease *	**Schizophrenia_1** *	**BMI** *	**Height** *	Ulcerative Colitis	**Cholesterol**	**Bipolar Disorder and Schizophrenia**	
Crohn's Disease	Ulcerative Colitis √	Rheumatoid Arthritis *	**Education Attainment - College** *	**BMI** *	**Triglycerides** *	**Parkinson**	Bipolar Disorder and Schizophrenia √	**Height** *
Breast Cancer_1	Prostate Cancer	**Fasting Glucose**	Lung Cancer	Upper Gastrointestinal Cancer	**Chronic Kidney Disease_2**	**Chronic Kidney Disease_1**	Acute Lymphoblastoid Leukemia	**AD_4**
Prostate Cancer	Breast Cancer_1	Lung Cancer	**Fasting Glucose**	**AD_4**	**Longevity_2**	Upper Gastrointestinal Cancer	**Parkinson**	**Coronary Artery Disease**
Lung Cancer	Prostate Cancer	**Fasting Glucose**	Breast Cancer_1	**Metformin**	Upper Gastrointestinal Cancer	**Longevity_2**		
Pancreatic Cancer	**Alzheimer_1**	**Bipolar Disorder and Schizophrenia**	**Crohn's Disease**	Glioma	Upper Gastrointestinal Cancer	**Glaucoma**	**ADHD**	**Juvenile Arthritis**
Fasting Glucose	HDL √	**Breast Cancer_1**	Triglycerides √	Diabetes_1 √	Fasting Insulin √	**Lung Cancer**	**Prostate Cancer**	Height *
Fasting Insulin	Fasting Glucose √	**Breast Cancer_1**	BMI	**Education Attainment - College**	Height	**HDL**	**Parkinson**	
Diabetes_1	Diabetes_2	Fasting Glucose √	Triglycerides √	HDL √	Height *	BMI √	**Chronic Kidney Disease_1**	LDL *
AD_3	AD_2	**Height** *	**BMI** *	AD_1	AD_4	Pubertal Growth Late Adolescence	Osteoporosis_1	**Coronary Artery Disease** *
Coronary Artery Disease	Fasting Glucose √	Diabetes_2 √	**AD_3** *	**Longevity_2**	**Neuroblastoma**	**Early Growth**	Smoking - Former Smoker √	Waist-Hip Ratio adjusted for BMI √

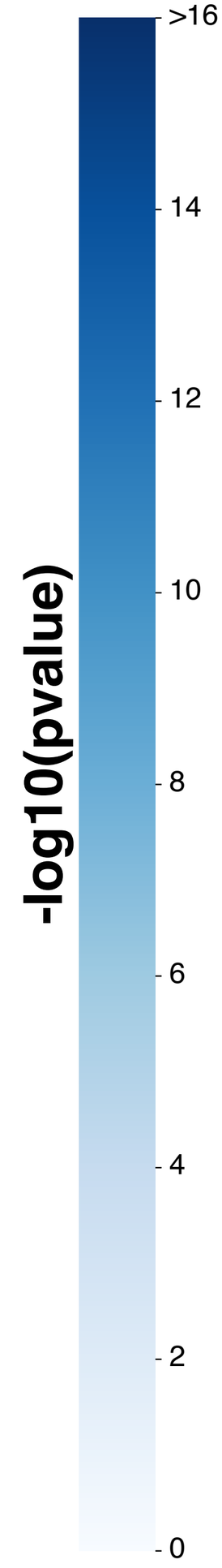


Figure 2

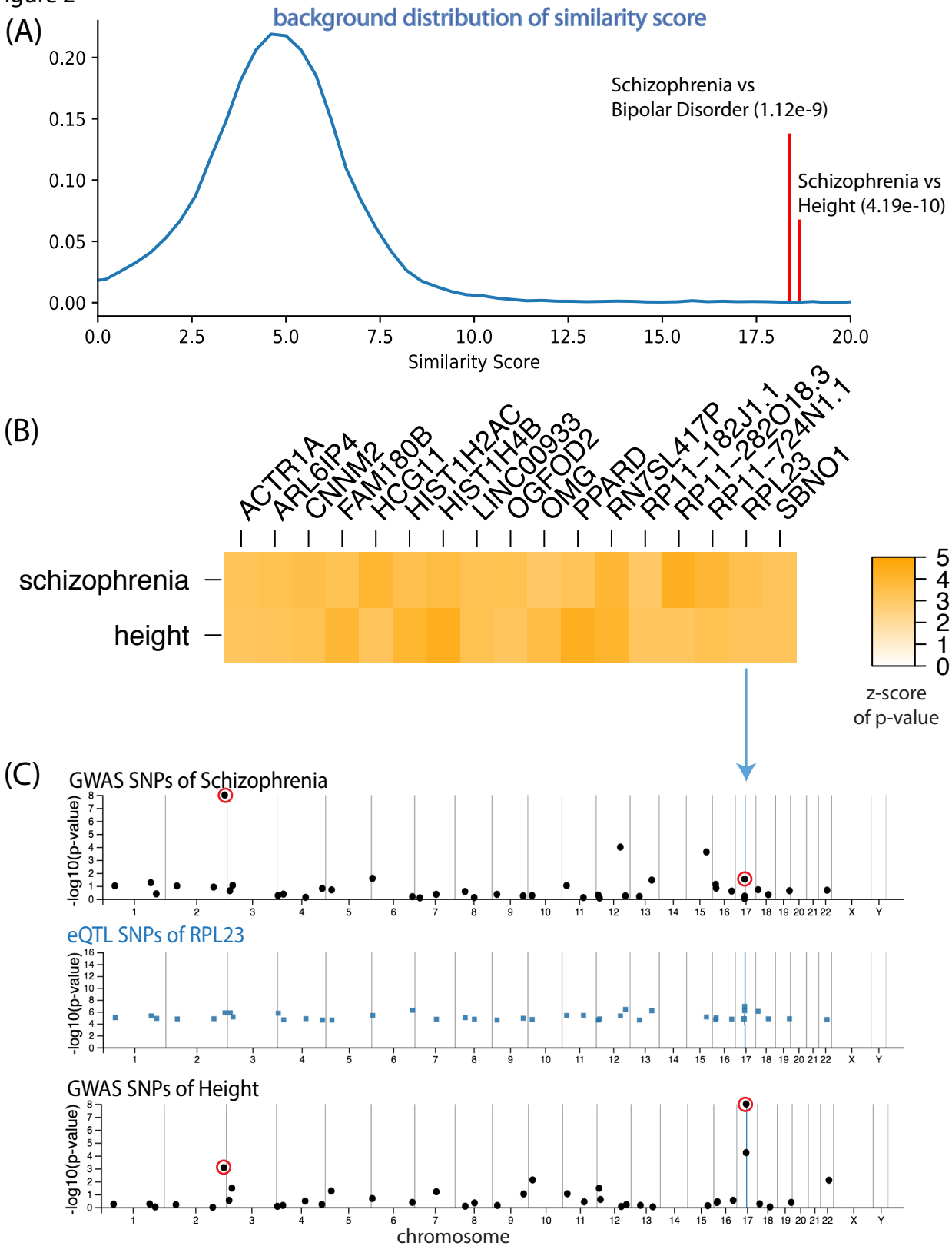

(A)
background distribution of similarity score
Schizophrenia vs
Bipolar Disorder (1.12e-9)
Schizophrenia vs
Height (4.19e-10)
Similarity Score
(B)
ACTR1A
ARL6IP4
CNNM2
FAM180B
HCG11
HIST1H2AC
HIST1H4B
LINC00933
OGFOD2
OMG
PPARD
RN7SL417P
RP11−182J1.1
RP11−282O18.3
RP11−724N1.1
RPL23
SBNO1
schizophrenia
height
z-score
of p-value
(C)
GWAS SNPs of Schizophrenia
eQTL SNPs of RPL23
GWAS SNPs of Height
-log10(p-value)
chromosome

Figure 3

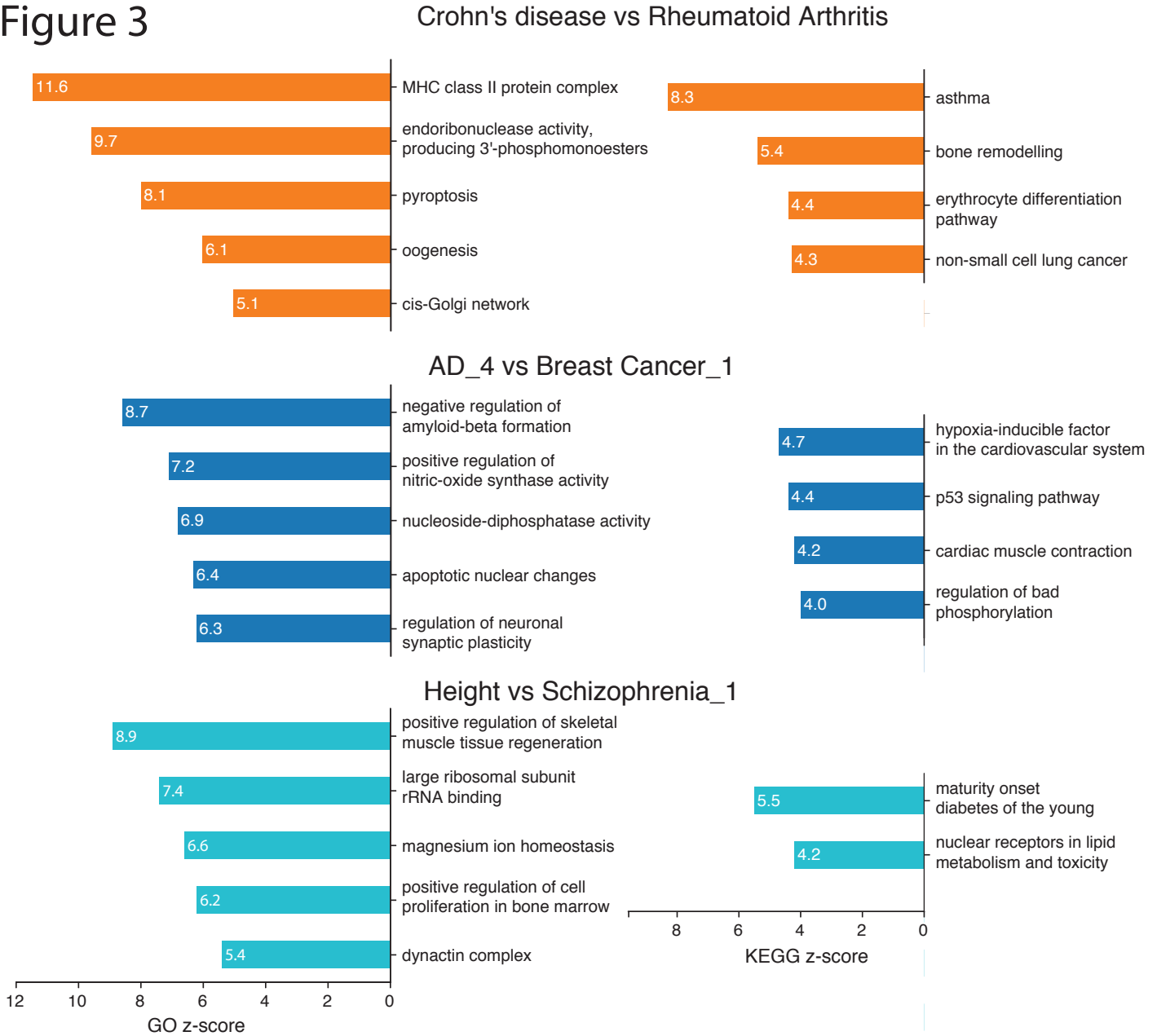
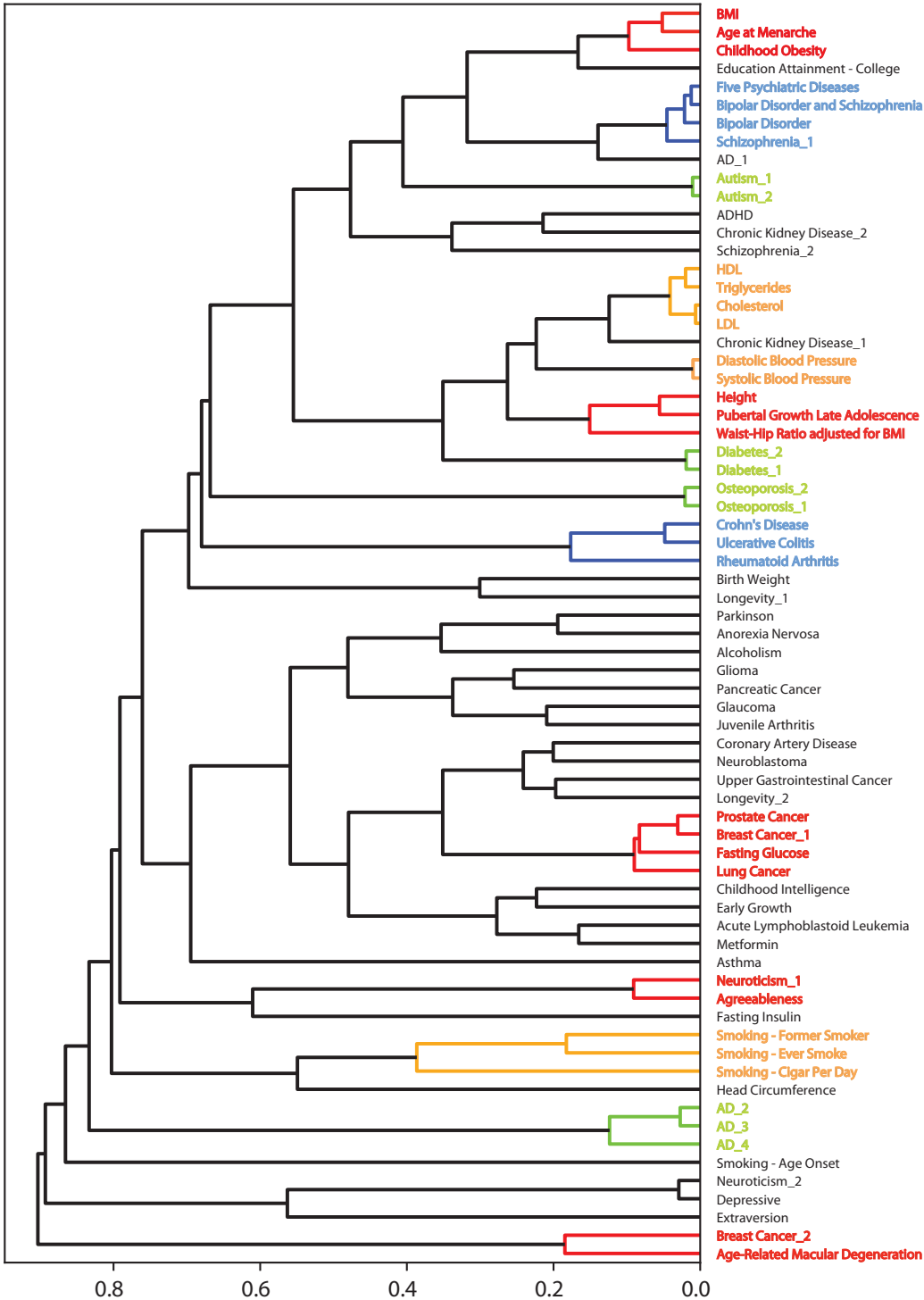

Crohn's disease vs Rheumatoid Arthritis
11.6
MHC class II protein complex
9.7
endoribonuclease activity, producing 3'-phosphomonoesters
8.1
pyroptosis
6.1
oogenesis
5.1
cis-Golgi network
8.3
asthma
5.4
bone remodelling
4.4
erythrocyte differentiation pathway
4.3
non-small cell lung cancer
AD_4 vs Breast Cancer_1
8.7
negative regulation of amyloid-beta formation
7.2
positive regulation of nitric-oxide synthase activity
6.9
nucleoside-diphosphatase activity
6.4
apoptotic nuclear changes
6.3
regulation of neuronal synaptic plasticity
4.7
hypoxia-inducible factor in the cardiovascular system
4.4
p53 signaling pathway
4.2
cardiac muscle contraction
4.0
regulation of bad phosphorylation
Height vs Schizophrenia_1
8.9
positive regulation of skeletal muscle tissue regeneration
7.4
large ribosomal subunit rRNA binding
6.6
magnesium ion homeostasis
6.2
positive regulation of cell proliferation in bone marrow
5.4
dynactin complex
12
10
8
6
4
2
0
GO z-score
5.5
maturity onset diabetes of the young
4.2
nuclear receptors in lipid metabolism and toxicity
8
6
4
2
0
KEGG z-score

Figure 4


BMI
Age at Menarche
Childhood Obesity
Education Attainment - College
Five Psychiatric Diseases
Bipolar Disorder and Schizophrenia
Bipolar Disorder
Schizophrenia_1
AD_1
Autism_1
Autism_2
ADHD
Chronic Kidney Disease_2
Schizophrenia_2
HDL
Triglycerides
Cholesterol
LDL
Chronic Kidney Disease_1
Diastolic Blood Pressure
Systolic Blood Pressure
Height
Pubertal Growth Late Adolescence
Waist-Hip Ratio adjusted for BMI
Diabetes_2
Diabetes_1
Osteoporosis_2
Osteoporosis_1
Crohn's Disease
Ulcerative Colitis
Rheumatoid Arthritis
Birth Weight
Longevity_1
Parkinson
Anorexia Nervosa
Alcoholism
Glioma
Pancreatic Cancer
Glaucoma
Juvenile Arthritis
Coronary Artery Disease
Neuroblastoma
Upper Gastrointestinal Cancer
Longevity_2
Prostate Cancer
Breast Cancer_1
Fasting Glucose
Lung Cancer
Childhood Intelligence
Early Growth
Acute Lymphoblastoid Leukemia
Metformin
Asthma
Neuroticism_1
Agreeableness
Fasting Insulin
Smoking - Former Smoker
Smoking - Ever Smoke
Smoking - Cigar Per Day
Head Circumference
AD_2
AD_3
AD_4
Smoking - Age Onset
Neuroticism_2
Depressive
Extraversion
Breast Cancer_2
Age-Related Macular Degeneration
0.8
0.6
0.4
0.2
0.0