Estimating Genetic Correlation from GWAS Summary Statistics

Brendan Bulik-Sullivan*, Hilary Finucane*, ... , *et. al.*

September 12, 2014

Contents

Abstract

Discovering relationships between phenotypes is a fundamental goal of epidemiology, with implications for drug development, nosology and treatment. Phenotypic correlations in observational epidemiological studies are often confounded by environmental factors, so genetic correlations between phenotypes may be more easily interpretable. The largest currently available sources of genotype-phenotype data are genome-wide association studies (GWAS); however, existing methods for estimating genetic correlation from GWAS data require genotype and phenotype data for at least one of the phenotypes, which is often impossible to obtain due to restrictions on data sharing. For this reason, only a few dozen genetic correlations have been estimated from GWAS data to date. In this paper, we describe a method based on LD Score regression which estimates genetic correlations directly from GWAS summary statistics and is immune to sample overlap. Since dozens of sets of summary statistics can be freely downloaded from the internet, we can report a much larger number of genetic correlations – more than 700 in this paper alone – than was previously possible. In addition, we relax many common assumptions about genetic architecture, and demonstrate that our method is not confounded when effect size depends on allele frequency or linkage disequilibrium.

1 Introduction

The (additive) genetic covariance, ρ_g between two phenotypes y_1 and y_2 is the bivariate analogue of heritability, and is defined as the covariance (in the population) between the additive genetic components of y_1 and y_2. The normalized version of genetic covariance is genetic correlation,

$$r_g := \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}, \tag{1.1}$$

which lies in the interval $[-1, 1]$. Note that genetic correlation is a stronger condition than pleiotropy. To exhibit genetic correlation, if is not sufficient for two phenotypes to be influenced by the same genetic loci: the directions of effect of the variants that influence the phenotypes must also be consistent across the genome.

Existing methods for estimating genetic correlation from genotype data (*e.g.,* restricted maximum likelihood (REML) as implemented in the software package GCTA [32, 33], or polgyenic risk scores [7]) require genotype data, which is often impossible to obtain due to restrictions on data sharing. For this reason, papers typically report at most a handful of genetic correlations, usually estimated from samples of at most a few tends of thousands of individuals, and only a few dozen genetic correlations have been estimated using GWAS data to date.

We propose a modification of LD Score regression [3] that can estimate genetic correlation from GWAS summary statistics. Precisely, if $z_{1,j} z_{2,j}$ are the Z-scores for a SNP j from two GWAS, and ℓ_j, the LD Score of SNP j, and we assume a simple model of genetic architecture where all SNP effect sizes are drawn *i.i.d.*, we can write

$$\mathbb{E}[z_{1,j} z_{2,j}] = \frac{\rho_g N_1 N_2}{M} \ell_j + \frac{N_s \rho}{N_1 N_2} \tag{1.2}$$

where N_1 and N_2 are the sample sizes of each study, N_s is the number of shared samples, ρ is the phenotypic covariance and ρ_g is the genetic covariance. This equation is derived in the supplementary note, and a similar relationship holds if one or both of the studies is an ascertained study of a binary phenotype. The relationship becomes more complicated if effect sizes are not

identically distributed, but instead depend on MAF or linkage disequilibrium; however, we can easily accommodate such dependencies with partitioned LD Score regression (as described in the results and methods section of this paper and [9]).

Thus, if we regress the product $z_{1,j}z_{2,j}$ of Z-scores from two GWAS against ℓ_j, the LD Score of SNP j, the slope times a constant estimates genetic covariance. Since sample overlap affects the term $z_{1,j}z_{2,j}$ equally for all SNPs, and the quantity N_s appears only as the intercept term, the LD Score regression estimator of genetic covariance is not biased by sample overlap[1]. We can estimate heritability using LD Score regression (as described in [3]), and use these heritability estimates to transform the estimates of genetic covariance into estimates of genetic correlation.

In this paper, we describe a series of simulations that validate these claims. We then replicate the genetic correlations reported by the PGC Cross-Disorder Group in [17], using only the summary statistics from [18], before reporting more than 700 (mostly novel) genetic correlations between phenotypes with publicly available GWAS summary data. We find that phenotypes generally tend to cluster within categories defined by clinical practice and observational epidemiology; nonetheless, we do observe some surprising results. For instance, we estimate genetic correlations close to zero between Alzheimers and schizophrenia, rheumatoid arthritis (RA) and schizophrenia, and smoking traits and schizophrenia, despite reports of psychosis in patients with Alzheimers, reduced risk of RA in patients with schizophrenia and high rate of smoking in patients with schizophrenia.

The computational demands of our method are mild, and we provide an open-source software package, `ldsc`, written in python, which implements the procedures described in this paper and also [3, 9] (URLs).

2 Results

2.1 Simulations

In order to check our derivations and verify the robustness of our inference procedure to violations of our modeling assumptions, we performed a variety of simulations.

2.1.1 Sample Overlap

To verify the unbiasedness of our estimation procedure in the presence of sample overlap (which is derived formally in the supplementary note), we simulated two GWAS with quantitative phenotypes, using genotypes from the 4,292 individuals in the Wellcome Trust Case/Control Consortium 1 (WTCCC1, [6]) bipolar disorder cohort for the first GWAS and genotypes from the 4,482 individuals in the WTCCC1 coronary artery disease cohort for the second GWAS. These cohorts contain 2,713 overlapping individuals. Additive genetic effect sizes were drawn from a bivariate point-normal distribution with 10% of SNPs causal and true genetic correlation 0.7. We then estimated genetic correlation using LD Score regression. Results from these simulations are summarized in [**SUPP TABLE N**], and confirm that LD Score regression is not confounded by sample overlap.

[1]Indeed, if ρ is known, for instance if both studies assay the same phenotype and $\rho = 1$, the intercept from this regression times a constant can be used as an estimator of the number of shared samples

2.1.2 Case-Control Ascertainment

We simulated ascertained GWAS in order to evaluate the performance of LD Score regression under various case/control ascertainment schemes. Simulating case/control GWAS under a liability threshold model requires rejection sampling from a large pool of genotypes. For instance, in order to simulate drawing 1,000 cases for a phenotype with prevalence of 1%, one would need on expectation to sample 100,000 genotypes. We do not have access to genotypes for 100,000 individuals, so we used simulated genotypes with a simplified LD structure ($r^2 = 0$ or 1) for simulating ascertainment.

We simulated standard case/control ascertainment following a liability threshold model, and estimated the genetic correlation using LD Score regression. Results from these simulations are summarized in [**SUPP TABLE N**], and confirm that LD Score regression recovers the true heritability and genetic correlation, even with heavy ascertainment. The simplified LD structure should not hinder interpretation of these simulation results, especially since we also provide a proof that LD Score regression is valid when applied to ascertained samples (Supplementary Note).

2.1.3 Complicated Ascertainment

Next, we simulated a more complicated ascertainment scheme, representative of the study design used by many large GWAS consortia, where all case samples are independent, but there is a large pool of healthy controls (*i.e.,* individuals who are controls for both phenotypes) shared between all studies. ... We caution that while LD Score regression estimates of genetic correlation are robust to standard case/control ascertainment and the healthy controls model of ascertainment, it can be difficult to interpret LD Score regression estimates of genetic covariance obtained from GWAS with more complicated ascertainment schemes. As an example, if one were to attempt to estimate the genetic correlation between body-mass index (BMI) and type-2 diabetes (T2D) using LD Score regression and summary statistics from a non-ascertained GWAS for BMI and a GWAS for T2D consisting of high-BMI controls and low-BMI cases, then the resulting estimate would not be on a readily-interpretable scale.

2.1.4 Misspecified Models of Genetic Architecture

Estimates of heritability and genetic covariance can be biased if the underlying model of genetic architecture is misspecified. For example, Speed, *et. al.* [27] demonstrate that GCTA-REML is confounded by MAF- or LD-dependent genetic architectures. Estimates of genetic correlation are somewhat more robust. Since genetic correlation is estimated as a ratio $\hat{\rho}_g / \sqrt{\hat{h}_1^2 \hat{h}_2^2}$ (or the weighted block jackknife estimator of this ratio, see Methods), model misspecification bias in the numerator and model misspecification bias in the the denominator will tend to approximately cancel, unless genetic correlation (not just heritability and genetic covariance) also depends on MAF or LD.

Naive LD Score regression is subject to similar biases as REML; however, it is possible to remove these biases by allowing for MAF- or LD-dependent genetic architectures by using partitioned LD Score regression (see [9] and Methods)

As a result, we expect simpler models tend to perform better as measured by mean square error (MSE). To test this hypothesis, we simulated a variety of realistic genetic architectures with MAF- and LD-dependence (WHICH MODELS?), and estimated genetic correlation using both naive LD Score regression and a more sophisticated inference procedure (MAF x LD-binned LD Score regression) that accounts for MAF and LD dependence. As expected, naive LD Score regression shows no

discernible directional bias and gives better MSE for genetic correlation estimation. Results from these simulations are summarized in supplementary table NNN. We note that this result holds only for genetic correlation, not for heritability or genetic covariance.

2.2 Real Data

2.2.1 Replication of PGC Cross Disorder Results

As a sanity check, we replicated the genetic correlation results obtained with raw genotypes and GCTA-REML in the PGC Cross-Disorder Group paper [17], using the summary statistics from [18] downloaded from the PGC website (URLs). For this replication, we used an LD Score with r^2's from the 1000 Genomes Europeans [4] but with the sum of r^2's taken only over SNPs in HapMap 3 [5] (hereafter referred to as HapMap3 LD Score) because this is most similar to the model of genetic architecture fit by GCTA-REML, where only the effects of genotyped SNPs are modeled (see the section "LD Score regression is Haseman-Elston Regression" in the Supplementary Note). The results from the PGC Cross-Disorder Group paper replicated closely, and the standard errors were similar to those obtained from GCTA-REML.

2.2.2 Application to a Large Set of Publicly Available Summary Statistics

Finally, we applied our method to a large set of publicly available summary statistics, including studies of schizophrenia [19], major depression [21], bipolar disorder [25], autism [18], ADHD [16], height [1], body mass index [28], waist-hip ratio [10], obesity [2], various insulin- and glucose-related traits [20, 24, 13, 30, 22, 8, 26], coronary artery disease [23], type-2 diabetes [15], rheumatoid arthritis [29], plasma lipid traits [31], inflammatory bowel disease [11] and Alzhiemer's disease [12].

A full list of studies, phenotypes and references is provided in supplementary table NNN. As a robustness check, we estimated genetic correlation using both naive LD Score regression and a more sophisticated model that allows for both MAF- and LD-dependent genetic architectures. As expected from our simulation results, results from both models are generally consistent (supplementary figure NNN). However, the simpler model gives much lower standard error when applied to real data, especially for smaller GWAS, and performed better in simulations as measured by MSE, we report results from the naive model hereafter (results from the more sophisticated model are displayed in Supplementary Figure NNN).

Note that LD Score regression heritability estimates are biased downwards by genomic control correction, so we cannot report heritability estimates for the phenotypes in our dataset.

3 Discussion

Recap of the highlights

Main point: it is now almost trivial (mod admixed or non european GWAS) to produce the all phenotype by all phenotype matrix of genetic correlations without the ethical issues around sharing genotypes IF people are willing to share INFO (or at least provide a file with QC+ INFO ¿ 0.9 SNPs)

4 Online Methods

4.1 Statistical Framework

See the supplementary note for a thorough exposition of the models behind LD Score regression.

4.2 Estimation of LD Scores

We estimated LD Scores from the European samples in the 1000 Genomes Project [4] reference panel using the --l2 flag in the ldsc software package by the authors (URLs) as in [3]. We estimated per-allele LD Scores using the --per-allele flag in ldsc, and we estimated MAF-binned LD Scores using the --cts-bin and --cts-breaks flags in ldsc. Following [3], we estimated LD Scores using a 1 centiMorgan (cM) window (with the ldsc flag --ld-wind-cm 1). Unlike [3], we used a MAF cutoff of 1% when estimating LD Scores, in order to reduce the impact of LD measurement error on our regressions. Since we only include variants with MAF $> 5\%$ in LD Score regressions for estimating genetic correlation, and because there is very little LD between variants with MAF $> 5\%$ and variants with MAF $< 5\%$, this is unlikely to impact our results. For the analyses with HapMap 3 [5] LD Scores, we took the sum of r^2's over the same subset of HapMap 3 SNPs retained for LD Score regression in [3], (that is, HapMap 3 SNPs with MAF $> 1\%$, excluding centromeres and regions with long-range LD) using the --keep flag in ldsc.

4.3 Quality Control

Imputation error can bias LD Score regression estimates of heritability and genetic covariance (Supplementary Note), though the biases in the numerator and denominator of the genetic correlation estimates will tend to cancel, so genetic correlation estimates are more robust to imputation error (Supplementary Figure NNN). To minimize this bias, we restricted to SNPs with reported INFO > 0.9 in all analyses, and we recommend restricting to INFO > 0.9 as best practice for all LD Score regressions.

Genomic control correction biases LD Score regression estimates of heritability and genetic covariance downwards (see the Supplementary Note of [3]); however, the bias in the numerator and denominator of the genetic correlation estimates will cancel, so genomic control correction will not bias LD Score regression estimates of genetic correlation. Nevertheless, we wished to obtain accurate heritability estimates as well as genetic correlation estimates, so we undid meta-analysis level genomic control correction by re-inflating all test statistics by multiplying by the reported λ_{GC} correction factor.

4.4 Partitioned LD Score Regression

In partitioned LD Score regression, we cut the set of SNPs in our reference panel into bins, for example, we might use five MAF bins, corresponding to MAF 0-10%, 10-20%, ... , 40-50%. We allow the variance explained per SNP to vary from bin to bin, but assume that variance explained per SNP is (roughly) equal within each bin. This amounts to approximating the unknown function that maps MAF to variance explained per SNP with a locally constant approximation. This presents a bias/variance tradeoff: if the mesh of our locally constant approximation is too coarse (*e.g.,* if we were to use two MAF bins instead of five), our locally constant approximation would be poor, and this would result in bias. On the other hand, if we use too many bins, the standard error of

our estimates will increase. However, we show in the simulations under the header "Misspecified Models of Genetic Architecture" that we can remove almost all MAF- and LD-bias under realistic parameter settings using only a few tens of bins, which increases the standard error only modestly.

4.4.1 Regression Weights

For heritability estimation, we use the LD Score regression weights derived in the supplementary note from [3]. The optimal regression weights for genetic covariance estimation are

$$\mathrm{Var}[\hat{\beta}_j\hat{\gamma}_j \,|\, \ell_j] = \left(\frac{h_1^2\ell_j}{M} + \frac{1-h_1^2}{N_1}\right)\left(\frac{h_2^2\ell_j}{M} + \frac{1-h_2^2}{N_2}\right) + 2\left(\frac{\rho_g\ell_j}{M} + \frac{\rho N_s}{N_1N_2}\right);$$

(Supplementary Note) however, this quantity depends on both heritabilities, the genetic covariance and the number of overlapping samples, which are often not known a priori, so some approximation is required. In order to obtain approximate regression weights, we use heritability estimates from the single-phenotype LD Score regressions, then we assume that N_s is close enough to zero that the term $\rho N_s/N_1N_2$ is negligible (though this default can be adjusted using the --overlap and --rho flags in ldsc), and estimate a rough genetic covariance (which we only use for the regression weights) using the aggregate estimator

$$\hat{\rho}_{g,agg} := \frac{1}{\bar{\ell}}\sum_{j=1}^{M}\hat{\beta}_j\hat{\gamma}_j,$$

where $\bar{\ell}$ denotes the mean LD Score among SNPs included in the regression. These regression weights are only an approximation to the optimal weights, but this will not introduce bias into the regression; it will only increase the standard error. The standard errors for LD Score regressions with summary statistics from GWAS with $N > 10,000$ are low enough to be interpretable, so non-optimality of the regression weights does not seem to be a major hindrance.

4.4.2 Genetic Correlation

Genetic correlation is defined as a ratio of quantities:

$$r_g := \frac{\rho_g}{\sqrt{h_1^2h_2^2}}.$$

The naive estimator of this ratio,

$$\hat{r}_{g,naive} := \frac{\hat{\rho}_g}{\sqrt{\hat{h}_1^2\hat{h}_2^2}},$$

is biased, and it is difficult to estimate a standard error for this estimator. Therefore, we use the block jackknife estimator of the ratio, which is less biased:

$$\hat{r}_{g,jackknife} := n_b\hat{r}_{g,naive} - \frac{n_b-1}{n_b}\sum_{i=1}^{n_b}\frac{\hat{\rho}_{g,i}}{\sqrt{\hat{h}_{1,i}^2\hat{h}_{2,i}^2}},$$

where n_b is the number of blocks, and $\hat{\rho}_{g,i}, \hat{h}_{1,i}^2$ and $\hat{h}_{2,i}^2$ are the estimates of genetic covariance and heritability obtained by deleting the i^{th} block. Our standard error estimates are also obtained from

the block jackknife:

$$\widehat{se}\left[\hat{r}_{g,jackknife}\right] := \sqrt{n_b}\sum_{i=1}^{n_b}\left(\frac{\hat{\rho}_{g,i}}{\sqrt{\hat{h}^2_{1,i}\hat{h}^2_{2,i}}} - \text{Mean}_i\left[\frac{\hat{\rho}_{g,i}}{\sqrt{\hat{h}^2_{1,i}\hat{h}^2_{2,i}}}\right]\right).$$

Block jackknife standard errors are robust to the correlated error structure of GWAS χ^2-statistics, so long as the block size exceeds the typical range of LD (see [3] and [14] for examples of papers in the statistical and population genetics literature that use this technique). We checked the reliability of our standard errors via simulations with real genotypes (Supplementary Table NNN), and found that the ldsc default setting of 2000 blocks genome-wide (which can be adjusted with the --num-blocks flag) gives standard error estimates that agree well with the empirical standard deviation across simulation replicates.

4.5 GWAS Data

4.5.1 Minimum Viable Summary Statistics

The minimum summary data required for estimating genetic correlation with LD Score regression are the following:

1. Genome-wide summary statistics from cohorts with similar ancestry
2. The summary statistics must be *signed* (allele and direction of effect)
3. The summary statistics should *never* be "corrected" via genomic control (GC) correction. Using GC'ed summary statistics will result in downward bias in the LD Score regression estimates of heritability and genetic covariance, and deflated LD Score regression intercepts, though the genetic correlation estimates will be fine.
4. The summary statistics must not be meta-analyzed with targeted genotyping at significant loci (*e.g.,* specialty genotyping arrays like immunochip, psych chip or metabochip, or replication cohorts)

The next details are nice to have, but are only used for filtering SNPs:

1. A measure of imputation quality (*e.g.,* INFO) for each SNP
2. Sample size at each SNP (for binary traits, number of cases and number of controls)
3. Sample MAF

If these data are not available, we recommend retaining only HapMap 3 SNPs with reference panel MAF above 5% for the LD Score regression as a workaround (note: for *regression*, not for estimation of the LD Scores), since HapMap3 SNPs seem to be well-imputed in most studies.

4.5.2 IGAP

International Genomics of Alzheimer's Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyze four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative, EADI; the Alzheimer Disease Genetics Consortium, ADGC; The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium, CHARGE; The Genetic and Environmental Risk in AD consortium, GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 and 2. Note that we only used stage 1 data for LD Score regression.

5 URLs

1. `ldsc` software:
   `github.com/bulik/ldsc`
2. LD block genotype simulation code:
   `github.com/bulik/ldsc-sim`
3. PGC (psychiatric) summary statistics:
   `www.med.unc.edu/pgc/downloads`
4. GIANT (anthopometric) summary statistics:
   `www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`
5. MAGIC (insulin, glucose) summary statistics:
   `www.magicinvestigators.org/downloads/`
6. CARDIoGRAM (coronary artery disease) summary statistics:
   `www.cardiogramplusc4d.org`
7. DIAGRAM (T2D) summary statistics:
   `www.diagram-consortium.org`
8. Rheumatoid Arthritis summary statistics:
   `www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/`
9. IGAP (Alzheimers) summary statistics:
   `www.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php`
10. IIBDGC (inflammatory bowel disease) summay statistics:
    `www.ibdgenetics.org/downloads.html`
    Note that we used a newer version of these data with 1000 Genomes imputation.
11. Plasma Lipid summary statistics:
    `www.broadinstitute.org/mpg/pubs/lipids2010/`
12. Coffee:
    `barismo.com`

References

[1] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[2] Sonja I Berndt, Stefan Gustafsson, Reedik Mägi, Andrea Ganna, Eleanor Wheeler, Mary F Feitosa, Anne E Justice, Keri L Monda, Damien C Croteau-Chonka, Felix R Day, et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature genetics*, 45(5):501–512, 2013.

[3] Brendan Bulik-Sullivan, Po-Ru Loh, Hilary Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *bioRxiv*, 2014.

[4] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[5] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[6] International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011.

[7] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

[8] Josée Dupuis, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, Anna L Gloyn, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116, 2010.

[9] Hilary K Finucane and Brendan Bulik-Sullivan. Partitioning heritability with ld score regression. *In preparation*, 2014.

[10] Iris M Heid, Anne U Jackson, Joshua C Randall, Thomas W Winkler, Lu Qi, Valgerdur Steinthorsdottir, Gudmar Thorleifsson, M Carola Zillikens, Elizabeth K Speliotes, Reedik Mägi, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature genetics*, 42(11):949–960, 2010.

[11] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[12] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 2013.

[13] Alisa K Manning, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, Ching-Ti Liu, Lawrence F Bielak, Inga Prokopenko, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics*, 44(6):659–669, 2012.

[14] Priya Moorjani, Nick Patterson, Joel N Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L Price, and David Reich. The history of african gene flow into southern europeans, levantines, and jews. *PLoS Genetics*, 7(4):e1001373, 2011.

[15] Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

[16] Benjamin M Neale, Sarah E Medland, Stephan Ripke, Philip Asherson, Barbara Franke, Klaus-Peter Lesch, Stephen V Faraone, Thuy Trang Nguyen, Helmut Schäfer, Peter Holmans, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(9):884–897, 2010.

[17] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature Genetics*, 2013.

[18] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381(9875):1371, 2013.

[19] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[20] Inga Prokopenko, Wenny Poon, Reedik Mägi, Rashmi Prasad, S Albert Salehi, Peter Almgren, Peter Osmark, Nabila Bouatia-Naji, Nils Wierup, Tove Fall, et al. A central role for grb10 in regulation of islet function in man. *PLoS genetics*, 10(4):e1004235, 2014.

[21] Stephan Ripke, Naomi R Wray, Cathryn M Lewis, Steven P Hamilton, Myrna M Weissman, Gerome Breen, Enda M Byrne, Douglas HR Blackwood, Dorret I Boomsma, Sven Cichon, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular psychiatry*, 18(4):497–511, 2012.

[22] Richa Saxena, Marie-France Hivert, Claudia Langenberg, Toshiko Tanaka, James S Pankow, Peter Vollenweider, Valeriya Lyssenko, Nabila Bouatia-Naji, Josée Dupuis, Anne U Jackson, et al. Genetic variation in gipr influences the glucose and insulin responses to an oral glucose challenge. *Nature genetics*, 42(2):142–148, 2010.

[23] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, 2011.

[24] Robert A Scott, Vasiliki Lagou, Ryan P Welch, Eleanor Wheeler, May E Montasser, Jian'an Luan, Reedik Mägi, Rona J Strawbridge, Emil Rehnberg, Stefan Gustafsson, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics*, 44(9):991–1005, 2012.

[25] Pamela Sklar, Stephan Ripke, Laura J Scott, Ole A Andreassen, Sven Cichon, Nick Craddock, Howard J Edenberg, John I Nurnberger, Marcella Rietschel, Douglas Blackwood, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nature genetics*, 43(10):977, 2011.

[26] Nicole Soranzo, Serena Sanna, Eleanor Wheeler, Christian Gieger, Dörte Radke, Josée Dupuis, Nabila Bouatia-Naji, Claudia Langenberg, Inga Prokopenko, Elliot Stolerman, et al. Common variants at 10 genomic loci influence hemoglobin a1c levels via glycemic and nonglycemic pathways. *Diabetes*, 59(12):3229–3239, 2010.

[27] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[28] Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Jian'an Luan, Reedik Mägi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42(11):937–948, 2010.

[29] Eli A Stahl, Soumya Raychaudhuri, Elaine F Remmers, Gang Xie, Stephen Eyre, Brian P Thomson, Yonghong Li, Fina AS Kurreeman, Alexandra Zhernakova, Anne Hinks, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature genetics*, 42(6):508–514, 2010.

[30] Rona J Strawbridge, Josée Dupuis, Inga Prokopenko, Adam Barker, Emma Ahlqvist, Denis Rybin, John R Petrie, Mary E Travers, Nabila Bouatia-Naji, Antigone S Dimas, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*, 60(10):2624–2634, 2011.

[31] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, Cristen J Willer, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[32] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

[33] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.