Genomic Analysis of Polygenic Traits

Bulik-Sullivan, B.K.

2016

document version
Publisher's PDF, also known as Version of record

Summary and General Discussion

In this thesis I describe a series of new statistical methods for learning about the biology and epidemiology of heritable traits disease using GWAS data. Large collaborative consortia have generated a massive quantity of genotype-phenotype data (see Chapter 2), but due to privacy concerns and data sharing restrictions, a majority of the individual-level data is siloed. It is not possible to work with more than a small fraction of the individual-level data, but almost all GWAS data can be used in the form of summary statistics. The best method for working with summary statistics will always be suboptimal relative to methods that work with individual-level data. Unfortunately, developing methods that work with summary statistics reduces the future incentive to share data. However, suboptimal methods that get the job done are preferable to optimal methods that cannot be applied. Therefore, the goals of this thesis are to improve both the quality of the GWAS summary statistics themselves (see Chapter 8) and the quality and range of inferences that one can make using these data.

What is the Origin of GWAS Inflation? (Chapter 3)

A trend observed across many large GWAS is that the genome-wide distribution of test statistics is inflated compared to the theoretical null [1, 2]. When we began investigating LD Score regression (circa mid-2013), the prevailing explanation was that GWAS inflation was the result of confounding due to inadequate control for population stratification or cryptic relatedness. An alternative explanation is that the inflation represented real polygenic signal, scattered across many variants with small effects [2]. Chapter 3 introduces a new statistical method called *LD Score Regression* that is able to distinguish between these alternatives. The basic idea is that under a polygenic model, inflation in χ^2 statistics will tend be higher in regions with high LD on average, but inflation from population stratification will be uncorrelated with LD. This observation gives us an observable that can be used to distinguish inflation due to confounding from inflation due to polygenicity. We applied this method to a large dataset of GWAS summary statistics. In all cases, the genomic control inflation factor λ_{GC} substantially overestimated the contribution of confounding to GWAS inflation, as measured by the LD Score regression intercept. Polygenicity accounts for the vast majority of GWAS inflation across all of the GWAS that we analyzed. In particular, we find that the contribution of population stratification to the results from the PGC schizophrenia GWAS [3] is very low.

This result has several important implications. First, these results justify the decision not to use genomic control correction in the PGC Schizophrenia GWAS and future GWAS. The value of λ_{GC} in that study was ≈ 1.4, so GC correction would have reduced the effective sample size by roughly 1.4-fold. Second, the LD Score regression intercept is a useful tool for calibrating association statistics. Loh *et al.* [4] describe a mixed model calibrated using LD Score regression. Fourth, the results of Chapter 3 demonstrate that genomic control correction is causing unnecessary loss of power, especially in large studies. For example, the LD Score regression intercept from the published GC-corrected summary statistics from [5] is 0.65, which means that the genomic control correction applied in that study reduced the effective sample size by 35%. In concrete terms, the number of individuals genotyped in that study was about 230,000, but the same number of GWAS hits could have been obtained using only about 150,000 individuals with no GC correction.

Combining Epigenetics and GWAS (Chapter 4)

In Chapter 4, we extend the LD Score Regression framework to partition heritability into components corresponding to various genomic annotations such as epigenetic marks. Chapter 4 addresses two main challenges. First, when quantifying the heritability accounted for by various functional categories using GWAS data, it is necessary to account for the fact that marginal regression coefficients are not estimates of causal SNP effect sizes. For example, introns are close to exons, so intronic SNPs will often be in LD with exonic SNPs. Coding regions are enriched for heritability [6], so intronic SNPs will often have large marginal regression coefficients due to LD with coding variants. If we were to naively assume that the marginal regression coefficients for intronic SNPs reflected the properties of intronic SNPs only, then we might mistakenly conclude that intronic regions were enriched for heritability. Second, naively partitioning SNPs into a small number of functional categories can result in omitted-variable bias. Suppose we wish to estimate the proportion of heritability accounted for by a regulatory category. One approach would be to partition the genome into two disjoint categories: SNPs in the category and SNPs not in the category. This will often be a poor model of genetic architecture, because regulatory elements tend to cluster near coding regions and coding regions are highly enriched for heritability. If we were to fit the naive model with only the category and its complement, we would mistakenly attribute a large proportion of the enrichment in heritability due to exons to the regulatory category.

In Chapter 4, we address these problems by fitting more flexible models of genetic architecture using a very large set of annotations. We demonstrate that using a large set of annotations gives the model sufficient flexibility to minimize omitted-variable bias. The downside of fitting a more complex model is an increase in variance; however, the ability of LD Score regression to operate on summary statistics allows us to use very large datasets that can handle a large number of covariates. We applied this method to 17 large GWAS. Our results include enrichment of heritability in conserved regions across many traits; immunological disease-specific enrichment of heritability in FANTOM5 enhancers [7]; and many cell-type-specific enrichments, including significant enrichment of central nervous system cell types in body mass index, age at menarche, educational attainment, and smoking behavior.

Using Genetics for High-Throughput Epidemiology (Chapter 5)

In Chapter 5, we develop a method for quantifying the genetic similarity between different traits that does not require genome-wide significant associations and takes as input GWAS summary statistics. Formally, our method estimates *genetic correlation*, which is a genome-wide measure of the shared heritability between two traits. The main challenge when estimating genetic

correlation from summary statistics is sample overlap. Many control cohorts are included in almost all GWAS datasets. Treating the summary statistics from one study as statistically independent from the summary statistics from all other studies would yield a false positive genetic correlation between any pair of case/control studies with shared controls or any overlapping pair of quantitative trait studies of correlated traits. Unfortunately, it is very difficult to determine the number of overlapping samples and the phenotypic correlation among the overlapping samples. Sample overlap data is stored in text in GWAS supplements, which are not machine-readable, and data about phenotypic correlation among overlapping samples is usually not reported anywhere. It was therefore necessary to develop a method to estimate sample overlap from summary statistics.

Somewhat surprisingly, under standard modeling assumptions about genetic architecture, sample overlap is identified from summary statistics alone and can be estimated using a generalization of LD Score regression. Under a polygenic model [8, 9], the expected value of the product of z-scores for a single SNP from two different studies, denoted $z_{1j}z_{2j}$ is

$$\mathbb{E}[z_{1j}z_{2j}] = \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j + \frac{\rho N_s}{\sqrt{N_1N_2}}, \tag{1}$$

where N_i is the sample size for study i, ρ_g is genetic covariance (defined in Methods), ℓ_j is LD Score [1], N_s is the number of individuals included in both studies, and ρ is the phenotypic correlation among the N_s overlapping samples. If study 1 and study 2 are the same study, then Equation 1 reduces to the single-trait result from [1] (Chapter 3), because genetic covariance between a trait and itself is heritability, and $\chi^2 = z^2$. This equation tells us that we can estimate genetic covariance using the slope from the linear regression of $z_{1j}z_{2j}$ against LD Score. If there is sample overlap, it will only affect the intercept from this regression (the term $\rho N_s/\sqrt{N_1N_2}$) and not the slope, so the estimates of genetic correlation will not be biased by sample overlap. Similar results hold if one or both studies is a case/control study, in which case genetic covariance is on the observed scale. This means that we can estimate genetic correlation using summary datasets with arbitrary, unknown sample overlap

We applied this estimator to all pairwise combinations of more than 35 traits. The results recapitulate many known epidemiological associations and non-associations, which we interpret as a validation of the method. There were several surprising results, the most intriguing of which is a positive genetic correlation between educational attainment and autism spectrum disorder. This result has since been replicated by other studies using different methods (polygenic scoring), independent datasets for educational attainment/cognitive function, and the same autism case/control dataset [10].

The Genome-Wide Contribution of Gene-by-Sex Interaction (Chapter 6)

In Chapter 6, we aimed to estimate the genome-wide contribution of (autosomal) gene-by-sex interactions to genetic architecture for several traits, especially sex-dimorphic traits. We used sex-specific summary statistics and cross-trait LD Score regression to estimate $r_{g,\mathsf{MF}}$ for 12 phenotypes. For all traits except waist-hip ratio, $r_{g,\mathsf{MF}}$ was in the range 0.85-1. This result suggests that the lack of discovery in previous scans for gene-by-sex interaction is not the result of low power; instead, the high estimates of $r_{g,\mathsf{MF}}$ indicate that scans for gene-by-sex interactions at larger sample sizes will yield few new discoveries.

LD Score and Haseman-Elston (Chapter 7)

In Chapter 7, I derive a connection between LD Score regression and existing estimators of heritability and genetic correlation. Precisely, I show that LD Score regression with constrained intercept, in-sample LD and inverse-LD Score weights is equivalent to Haseman-Elston (HE) regression [11, 12]. This result holds also if fixed-effect covariates (*e.g.,* principal components of the genotype matrix [13, 14]) are included in the model. LD Score regression with default weights is slightly more efficient than HE regression. For non-ascertained studies of quantitative traits, both HE regression and LD Score regression are less efficient estimators of heritability and genetic correlation than restricted maximum likelihood (REML). For ascertained studies of case/control traits, REML yields heritability estimates that are biased downwards, and HE regression is the state-of-the-art SNP-heritability estimator [15, 16]. Chapter 7 shows that it is possible to obtain slightly more efficient estimates of heritability and genetic covariance for ascertained studies of binary traits using LD Score regression, with the additional advantage that LD Score regression is much less computationally expensive than HE regression.

Mixed Models for Meta-Analysis and Sequencing (Chapter 8)

In Chapter 8, I show that meta-analyses of mixed model association test statistics typically achieve much lower power than a mixed model association statistic computed on the entire pooled dataset. However, computing a mixed model association test statistic on the pooled dataset requires sharing all individual-level genotype and phenotype data with a centralized meta-analysis group, which is often not possible.

The mixed model score test from BOLT-LMM [17] is simply a regression on prediction residuals. The gain in power compared to linear regression depends on the prediction R^2. In a meta-analysis of mixed models, we train one predictor for each cohort using only the training data from that cohort, which results in poor predictions and little gain in power, because cohorts tend to be small. In contrast, if we perform mixed model association testing on the pooled dataset, we train one predictor on the entire dataset, which results in good predictions and considerable gain in power, because the pooled dataset is large. The key observation in Chapter 8 is that there is no requirement that the predictions be obtained using in-sample penalized linear regression. To maximize power, we should use whatever procedure we can to maximize the prediction R^2 (taking care to avoid proximal contamination [18]). In particular, we can use predictions obtained using out-of-sample data, including predictors that require only summary data and a small validation set for training (*e.g.,* polygenic scoring [19] or LDpred [20]).

The technique that I recommend in Chapter 8 is for meta-analysis consortia to perform GWAS using BOLT-LMM [17] on each cohort using a polygenic score generated from whole-consortium summary statistics (leaving one chromosome out at a time

to avoid proximal contamination) as a covariate. Using BOLT-LMM corrects for population stratification and relatedness, and using the polygenic score as a covariate increases power by de-noising the phenotype [18]. This approach will have intermediate power between a meta-analysis of mixed models and a mixed model on the pooled dataset; the relative power of the three approaches depends on the average cohort size (many small cohorts means that a meta-analysis of mixed models will perform poorly). The potential gain in power compared to linear regression with principal components or a meta-analysis of mixed models could be quite large: based off published polygenic score results, [5, 21] I estimate that the next height GWAS could increase effective sample size by ∼15% and the next BMI GWAS could increase effective sample size by ∼8%.

General Conclusion

We have described a series of methods for asking questions about the biology and etiology of human traits and diseases using GWAS data, with a focus on methods that can be applied to large GWAS meta-analyses that do not share individual-level data. The key theme of these methods is that analyses of GWAS data benefit from taking into account the high-dimensional nature of the data, both in phenotype and variant space (*i.e.,* high dimensionality both on the left and right side of the regression equation). Joint modeling of many phenotypes and many variants offers many advantages over the traditional approach of considering only marginal single-variant single-phenotype associations.

1 References

[1] 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. *Nature Genetics*, 2015.

[2] Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, Massimo Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

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

[4] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, et al. Efficient bayesian mixed model analysis increases association power in large cohorts. *bioRxiv*, page 007799, 2014.

[5] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[6] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[7] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014.

[8] 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.

[9] Sang Hong Lee, Jian Yang, Michael E Goddard, Peter M Visscher, and Naomi R Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542, 2012.

[10] TK Clarke, MK Lupton, AM Fernandez-Pujals, J Starr, G Davies, S Cox, A Pattie, DC Liewald, LS Hall, DJ MacIntyre, et al. Common polygenic risk for autism spectrum disorder (asd) is associated with cognitive ability in the general population. *Molecular psychiatry*, 2015.

[11] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2(1):3–19, 1972.

[12] Robert C Elston, Sarah Buxbaum, Kevin B Jacobs, and Jane M Olson. Haseman and elston revisited. *Genetic epidemiology*, 19(1):1–17, 2000.

[13] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[14] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[15] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[16] Guo-Bo Chen. Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman–elston regression. *Frontiers in genetics*, 5, 2014.

[17] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 2015.

[18] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–106, 2014.

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

[20] Bjarni Vilhjalmsson, Jian Yang, Hilary Kiyo Finucane, Alexander Gusev, Sara Lindstrom, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *bioRxiv*, page 015859, 2015.

[21] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 2014.