Mixed Models for Meta-Analysis and Sequencing

Brendan Bulik-Sullivan

May 29, 2015

Abstract

Mixed models are an effective statistical method for increasing power and avoiding confounding in genetic association studies. Existing mixed model methods have been designed for "pooled" studies where all individual-level genotype and phenotype data are simultaneously visible to a single analyst. Many studies follow a "meta-analysis" design, wherein a large number of independent cohorts share only summary statistics with a central meta-analysis group, and no one person can view individual-level data for more than a small fraction of the total sample. When using linear regression for GWAS, there is no difference in power between pooled studies and meta-analyses [1]; however, we show that when using mixed models, standard meta-analysis is much less powerful than mixed model association on a pooled study of equal size. We describe a method that allows meta-analyses to capture almost all of the power available to mixed model association on a pooled study without sharing individual-level genotype data. The added computational cost and analytical complexity of this method is minimal, but the increase in power can be large: based on the predictive performance of polygenic scoring reported in [2] and [3], we estimate that the next height and BMI studies could see increases in effective sample size of ≈15% and ≈8%, respectively. Last, we describe how a related technique can be used to increase power in sequencing, targeted sequencing and exome array studies.

Note that these techniques are presently only applicable to randomly ascertained studies and will sometimes result in loss of power in ascertained case/control studies. We are developing similar methods for case/control studies, but this is more complicated.

Notation

We use the following notation throughout:

- $N \in \mathbb{N}$: sample size.
- $y \in \mathbb{R}^N$: centered and standardized phenotype.
- $X_j \in \mathbb{R}^N$: centered and standardized vector of genotypes at the test variant j.
- $\beta_j \in \mathbb{R}$: effect size of the test variant j.
- $\boldsymbol{X}_{-j}$: $M \times N$ matrix of centered standardized genotypes excluding a region around j.

As in [4], we assume that all fixed effects (*e.g.,* age, sex, batch, study indicators, principal components of the genotype matrix *etc*) have already been projected out of the data. The meta-analysis method proposed here achieves protection from population stratification and proper calibration with related samples using BOLT-LMM. In order to simplify notation, we omit most discussion of

population stratification and family structure; these issues have already been addressed with regard to BOLT-LMM in ref [4].

Our goal is to test the null hypothesis $\mathcal{H}_0 \colon \beta_j = 0$.

Association Testing with Linear Regression

The simplest approach to testing genetic variants for association with a quantitative phenotype is to fit the model

$$\underbrace{y}_{\text{Phenotype}} = \underbrace{X_j \beta_j}_{\text{Test Variant}} + \underbrace{\epsilon}_{\text{Noise}}. \tag{1}$$

using linear regression (typically we would also include the top several principal components of the genotype matrix as covariates; as previously noted, we assume without loss of generality that these have already been projected out). Linear regression yields an effect-size estimate $\beta_{j,\text{lin}}$ and a variance estimate $\hat{\sigma}^2_{j,\text{lin}}$ We can construct χ^2 statistics

$$\chi^2_{\text{lin}} \coloneqq \frac{\hat{\beta}^2_{j,\text{lin}}}{\hat{\sigma}^2_{j,\text{lin}}}. \tag{2}$$

This test statistic is called the Armitage Trend Test (ATT) statistic. The choice of linear regression statistic is not particularly important; all of the standard linear regression test statistics are asymptotically equivalent. The ATT statistic asymptotically follows a one degree-of-freedom noncentral χ^2 distribution. For small β_j (typical in GWAS), $\sigma^2_{j,\text{lin}} \coloneqq \text{Var}[\hat{\beta}_{j,\text{lin}}] \approx 1/N$, so the noncentrality parameter is

$$\text{NCP}_{\text{lin}} = N\beta_j^2. \tag{3}$$

Note that since β_j is the standardized effect size of variant j, β_j^2 is variance explained.

Association Testing with Mixed Models

The linear regression model from Equation 1 treats the effects of all variants other than the test variant as noise. We can increase the power of the test for association by explicitly modeling the effects of the other variants. The model then becomes

$$\underbrace{y}_{\text{Phenotype}} = \underbrace{X_j \beta_j}_{\text{Test Variant}} + \underbrace{f(\boldsymbol{X}_{-j})}_{\text{Other Variants}} + \underbrace{\epsilon'}_{\text{Noise}}. \tag{4}$$

Note that the noise term ϵ' in Equation 4 has smaller variance than the noise term ϵ in Equation 1, because $\epsilon = \epsilon' + f(\boldsymbol{X}_{-j})$. If we allow f to take arbitrary functional form, then this is *additive* mixed model; a linear mixed model is the special case where we constrain f to be a linear function. This model is fit using a two-step procedure [4]:

1. Generate predictions $\hat{f}(\boldsymbol{X}_{-j})$ and prediction residuals $y_{\text{resid}} \coloneqq y - \hat{f}(\boldsymbol{X}_{-j})$.
2. Obtain an effect-size estimate and variance $\hat{\beta}_j$ and $\hat{\sigma}^2_j$ by fitting the model $y_{\text{resid}} = X_j \beta_j + \epsilon$ using linear regression.

Refs [4, 5] refer to step 1 as 'de-noising' the phenotype. It is important to use predictions generated with the test variant and variants in LD with the test variant left out. By including the effect of the test variant in the predictions, we remove most of the effect of the test variant from the residual phenotype, which results in loss of power in step 2. Ref [6] coined the term 'proximal contamination' to describe this problem. We can avoid proximal contamination by generating predictions with the test variant and its LD partners removed (for example, ref [4] train 22 predictors, leaving out one chromosome at a time).

The mixed model fitting algorithm can be used with arbitrary choice of predictor $\hat{f}$. A large number of different mixed model association methods have been published in the statistical genetics literature, and the distinctions between most of these methods can be understood as different statistical and algorithmic choices for fitting $\hat{f}$ (at least in the special case where individuals in the study are unrelated). For example, GCTA generates predictions using BLUP [5], which is equivalent to ridge regression with the shrinkage parameter set via maximum-likelihood. FAST-LMM [7] uses faster algorithms to fit the same model (GEMMA [8] and EMMAX [9] also fit the same model, but without avoiding proximal contamination and so suffer from loss of power [5]). FAST-LMM-Select [6] uses ridge regression with a feature selection step. LMM-LASSO [10] uses a Lasso with shrinkage parameter selected via cross-validation. BOLT-LMM [4] uses Bayesian linear regression with a mixture-of-Gaussians prior with hyperparameters selected via a combination of (stochastic) maximum-likelihood and cross-validation. This list of examples is not intended to be exhaustive, but illustrates the general principle.

The variance of the standardized effect size estimate $\hat{\beta}_{j,\mathrm{lmm}}$ from the mixed model applied to a single (pooled) study is $\sigma^2_{j,\mathrm{lmm}} \coloneqq \mathrm{Var}[\hat{\beta}_{j,\mathrm{lmm}}] = (1 - R^2_{-j})/N$, where $R^2_{-j} \coloneqq \mathrm{Cor}[f(\boldsymbol{X}_{-j}), \hat{f}(\boldsymbol{X}_{-j})]^2$ is the prediction R^2 achieved by the mixed model predictor (leaving out a region around variant j in order to avoid proximal contamination). Thus, we can construct mixed-model χ^2 statistics

$$\chi^2_{\mathrm{lmm}} \coloneqq \frac{\hat{\beta}^2_{j,\mathrm{lmm}}}{\hat{\sigma}^2_{j,\mathrm{lmm}}}, \tag{5}$$

which asymptotically follow a noncentral one degree-of-freedom χ^2 distribution with noncentrality parameter

$$\begin{aligned} \mathrm{NCP}_{\mathrm{lmm}} &= \frac{N\beta^2_j}{1 - R^2_{-j}} \\ &= \frac{\mathrm{NCP}_{\mathrm{lin}}}{1 - R^2_{-j}}. \end{aligned} \tag{6}$$

Note that the power increase depends on the *true* prediction R^2, *i.e.,* the squared correlation between $\hat{f}(\boldsymbol{X}_{-j})$ and $f(\boldsymbol{X}_{-j})$ (which is not observable) [4]. The training sample prediction R^2, *i.e.,* $\widehat{\mathrm{Cor}}[y, \hat{f}(\boldsymbol{X}_{-j})]^2$, will typically over-estimate the true prediction R^2, because error metrics on the training set are optimistic. For power calculations, one can obtain a better estimate of R^2_{-j} using cross-validation, as in [4].

The χ^2 statistic from Equation 5 is equivalent to the retrospective quasilikelihood score statistic from BOLT-LMM in the case where $\mathrm{Var}[X_j] = \boldsymbol{I}$, *i.e.,* for GWAS that sample unrelated individuals (ref [4], Supplementary Equation 22). Association testing in GWAS datasets with family relatedness is more difficult; in particular, residualizing on predictions generated from typed SNPs (equivalently,

using the genetic relatedness matrix estimated from typed SNPs) achieves suboptimal power and suffers from inflated type I error due to residual non-independence of family members (*e.g.,* from correlated environmental effects, or the effects of untyped rare variants). For an overview of mixed model methods applied to family data, see ref [11].

Meta-Analysis Terminology and Study Design

For readers unfamiliar with the analysis procedure used by consortia such as GIANT, we provide a brief review of GWAS meta-analysis.

Gathering large samples of genotype-phenotype data is difficult, time-consuming and expensive. As a result, most GWAS datasets are generated not by a single lab, but rather by large consortia that pool the efforts of many research groups working in parallel. We refer to the data generated by an individual research group as a 'cohort'.

One approach to GWAS is to aggregate the genotype and phenotype data from all cohorts onto a single computer, then to run regressions on the combined dataset. We refer to this as a 'pooled' study design. Some examples of consortia that use the pooled design include the Psychiatric Genomics Consortium [12–16] and the International Inflammatory Bowel Disease Consortium [17].

Studies that use the pooled design represent a minority of all GWAS, because there are often restrictions (imposed by national law, IRB regulations, etc) that prohibit researchers from sharing individual-level genotype data with other groups. When researchers cannot share individual-level genotype data, an alternative approach is for each research group to run regressions within their own cohort, then for the research groups to share summary statistics (effect size estimates and variances, or equivalent) with a central meta-analysis group, who then meta-analyze the summary statistics using the methods described in the next section. We refer to this study design as a 'meta-analysis'.

Inverse-Variance Meta-Analysis

Suppose we have summary-level association results from S non-overlapping studies of the same phenotype, where the summary data consist of a (standardized) effect-size estimate $\hat{\beta}_{jk}$ and a variance estimate $\hat{\sigma}^2_{jk}$ for $k = 1, \ldots, S$. We suppose that the $\hat{\beta}_{jk}$ are consistent estimates of a single underlying parameter β_j, and that the $\hat{\sigma}^2_{jk}$ are consistent estimates of the true sampling variances σ^2_{jk}. For asymptotically normal estimators (a class which includes all both the linear regression and mixed model estimators described here), the sampling distribution is $\hat{\beta}_{jk} \sim \mathcal{N}(\beta_j, \sigma^2_{jk})$. In GWAS, the standard approach for combining these data is inverse-variance meta-analysis [18], which is the minimum-variance unbiased estimator of the true effect β_j The inverse variance meta-analysis effect size and variance estimates are

$$\hat{\beta}_{j,\text{meta}} := \frac{\sum_{k=1}^{S} \hat{\beta}_{jk}/\hat{\sigma}^2_{jk}}{\sum_{k=1}^{S} 1/\hat{\sigma}^2_{jk}}; \tag{7}$$

$$\hat{\sigma}^2_{j,\text{meta}} := \frac{1}{\sum_{k=1}^{S} 1/\hat{\sigma}^2_{jk}}. \tag{8}$$

The asymptotic sampling distribution is therefore $\beta_{j,\text{meta}} \sim \mathcal{N}(\beta_j, \sigma^2_{j,\text{meta}})$. The meta-analysis χ^2 statistic is

$$\chi^2_{j,\text{meta}} \coloneqq \frac{\hat{\beta}^2_{j,\text{meta}}}{\hat{\sigma}^2_{j,\text{meta}}},$$

which follows a one degree-of-freedom noncentral χ^2 distribution with noncentrality parameter

$$\text{NCP}_{\text{meta}} = \frac{\beta^2_j}{\sigma^2_{j,\text{meta}}}.$$

Here, $\sigma^2_{j,\text{meta}} \coloneqq \left(\sum_{k=1}^{S} 1/\sigma^2_{jk}\right)^{-1}$ is the true (rather than estimated) meta-analysis variance.

Meta-Analysis Noncentrality Parameters

The noncentrality parameter of an inverse-variance weighted meta-analysis of linear regression association tests is

$$\begin{aligned}\text{NCP}_{\text{lin,meta}} &= \beta^2_j \sum_{k=1}^{S} N_k \\ &= N_{\text{meta}} \beta^2_j \\ &= \text{NCP}_{\text{lin,pool}},\end{aligned} \tag{9}$$

where $N_{\text{meta}} \coloneqq \sum_k N_k$, and $\text{NCP}_{\text{lin,pool}}$ denotes the noncentrality parameter of linear regression applied to a pooled study with sample size N_{meta}. Hence, when using linear regression for association testing, there is no difference in power between a pooled study and a meta-analysis of equal size [**?**].

The noncentrality parameter of an inverse-variance weighted meta-analysis of mixed model association tests is

$$\text{NCP}_{\text{lmm,meta}} = \beta^2_j \sum_{k=1}^{S} \frac{N_k}{1 - R^2_{-j,k}},$$

where $R^2_{-j,k}$ denotes the prediction R^2 achieved by the mixed model predictor leaving a region around variant j out trained using only the samples from study k. The previous expression is exact, and should be preferred for power calculations. In order to obtain a simpler expression for intuition, we make the approximation $1/(1 - R^2) \approx R^2 + 1$ (which holds for small R^2). This allows us to (approximately) simplify the previous expression to

$$\begin{aligned}\text{NCP}_{\text{lmm,meta}} &\approx \beta^2_j \sum_{k=1}^{S} N_k (R^2_{-j,k} + 1) \\ &= N_{\text{meta}} (\bar{R}^2_{-j} + 1),\end{aligned} \tag{10}$$

where

$$\bar{R}^2_{-j} \coloneqq \frac{1}{N_{\text{meta}}} \sum_{k=1}^{S} N_k R^2_{-j,k}$$

is the sample-size weighted average prediction R^2 achieved by mixed model predictors trained using data from one cohort at a time (leaving out a region around the test variant j in order to avoid proximal contamination). In contrast, the noncentrality parameter of mixed model association applied to a pooled study of equal size is

$$\mathrm{NCP}_{\mathrm{lmm,pool}} = \frac{\mathrm{NCP}_{\mathrm{lin}}}{1 - R^2_{-j,\mathrm{pool}}},$$

where $R^2_{-j,\mathrm{pool}}$ denotes the prediction R^2 achieved by the mixed model predictor trained on the pooled sample (again, leaving out the region around variant j). In a typical application, N_{meta} might be on the order of hundreds of thousands, while N_i will be one to two orders of magnitude smaller. Since a predictor trained on hundreds of thousands of individuals will perform much better than a predictor trained on tens of thousands of individuals, $R^2_{\mathrm{pool}} > \bar{R}^2$. Thus, the relationship between the four noncentrality parameters described above is

$$\mathrm{NCP}_{\mathrm{lmm,pool}} > \mathrm{NCP}_{\mathrm{lmm,meta}} > \mathrm{NCP}_{\mathrm{lin,meta}} = \mathrm{NCP}_{\mathrm{lin,pool}}. \tag{11}$$

More Powerful Strategies for Mixed Model Meta-Analysis

The origin of the loss of power in mixed model meta-analysis compared to pooled mixed model association is the poor performance of predictors trained only on individual cohorts. Therefore, what is required is a method for training a predictor on the whole dataset without sharing genotype data; that is, a method for training a predictor using the summary statistics. Conveniently, such methods have already been described [19–21]. At the time of writing, the most sophisticated general-purpose method in this class appears to be LDpred [19]; though we emphasize that this meta-analysis procedure is modular with regard to choice of predictor, so we can substitute more sophisticated predictors as they become available or as is appropriate for particular phenotypes.

In order to achieve maximum power gain without sharing genotype data, a meta-analysis consortium would need to employ a iterative computational scheme[1], which would require multiple exchanges of summary data between the cohorts and the meta-analysis group. Coordinating the submission of summary statistics to the central meta-analysis group is typically a bottleneck, so iterative approaches that require multiple rounds of data exchange will take much longer than a standard meta-analysis. Therefore, in this section we describe a simple one-step approach that provides an attractive compromise between speed and power.

The idea behind the fast, one-step design is that the meta-analysis consortium could construct a predictor using the summary statistics from the *last* meta-analysis. For example, the 2014 height study [2] could have used a predictor trained on the summary statistics from the 2010 height study [22], and the next height GWAS could use a predictor constructed from the 2014 height summary statistics. Similarly, the 2015 BMI study [3] could have used a predictor trained on the 2010 BMI study [23], and so forth.

The noncentrality parameter for this strategy is

$$\mathrm{NCP}_{\mathrm{summary}} = \frac{\mathrm{NCP}_{\mathrm{lin}}}{1 - R^2_{-j,\mathrm{summary}}}, \tag{12}$$

[1]There are algorithms from machine learning that allow one to train a predictor by computing on only a subset of the data at any given time (*e.g.,* stochastic gradient descent and its more sophisticated parallel cousins). In machine learning, the barrier to holding all data in memory is typically data size, rather than privacy concerns, but the algorithms could nevertheless be adapted to GWAS consortia.

where $R^2_{-j,\text{summary}}$ denotes the prediction R^2 achieved by the predictor trained on summary statistics (leaving out the region around j).

There is no firm relationship between $R^2_{-j,\text{summary}}$ and $R^2_{-j,\text{pool}}$. Our intuition is that when comparing equally sophisticated prediction algorithms, $R^2_{-j,\text{summary}}$ will typically be slightly lower than $R^2_{-j,\text{pool}}$, (which implies $\text{NCP}_{\text{summary}} < \text{NCP}_{\text{pool}}$) because some information is lost when working with summary data instead and reference LD matrices instead of individual-level data and sample LD matrices [19].

Nevertheless, in typical scenarios where each N_i is one to two orders of magnitude smaller than N_{meta}, we expect that the difference between $R^2_{-j,\text{pool}}$ and $R^2_{-j,\text{summary}}$ will be much less than the difference between $R^2_{-j,\text{summary}}$ and $\bar{R}^2$, because both $R^2_{-j,\text{pool}}$ and $R^2_{-j,\text{summary}}$ reflect the performance of predictors trained on a training set much larger than the training sets available to the predictors whose performance is measured by $\bar{R}^2_{-j}$.

Thus, the relationship between the noncentrality parameters is

$$\text{NCP}_{\text{lmm,pool}} \gtrsim \text{NCP}_{\text{summary}} > \text{NCP}_{\text{lmm,meta}} > \text{NCP}_{\text{lin,meta}} = \text{NCP}_{\text{lin,pool}}. \tag{13}$$

Mixed Models for Sequencing Studies

The key idea behind the procedure in the previous section can be summarized in the following way: the predictions used in mixed model association testing do not need to be trained on the same sample used for testing. We should use the best genetic predictor (or combination of predictors) available, no matter how it is trained. This idea is also applicable to sequencing studies. At the time of writing, sequencing is roughly an order of magnitude more expensive than array genotyping; consequently, sequenced datasets tend to be roughly an order of magnitude smaller than genotyped datasets for the same trait. If we simply applied a mixed model (*e.g.,* BOLT-LMM) to the smaller sequenced datasets, the prediction R^2 achieved by the mixed model predictor would be poor compared to what we could achieve by training a predictor on the larger genotyped datasets. Thus, we can increase power by using a predictor trained on the largest available genotyped dataset when running mixed-model association testing in the sequence data. Precisely,

1. Train a predictor on the SNP-array GWAS data. If individual-level genotype data are available, we can use the `--predBetasFile` flag in `bolt` to export prediction weights. If not, we can use LDpred to convert summary statistics into valid prediction weights.

2. Evaluate this predictor on the sequenced individuals (*e.g.,* using the `--score` flag in `plink` [24, 25]).

3. Perform association testing on the sequenced dataset using `bolt` with the predictions from the previous step as covariates *e.g.,* using the `--covarFile` and `--qCovarCol` flags (again, leaving one chromosome out at a time in order to avoid proximal contamination).

The noncentrality parameter for this test is the same as in Equation 12; the gain in power vs standard in-sample mixed model association depends on the increase in prediction R^2 gained by training a predictor on the SNP-array GWAS data. Note that the p-values reported by `bolt` are only asymptotically valid, so the calibration may be poor for variants with very low minor allele count.

1 References

[1] DY Lin and D Zeng. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic epidemiology*, 34(1):60–66, 2010.

[2] 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.

[3] 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.

[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. *Nature Genetics*, 2015.

[5] 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.

[6] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6):525–526, 2012.

[7] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[8] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[9] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[10] Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, 2013.

[11] George Tucker, Po-Ru Loh, Iona M MacLeod, Ben J Hayes, Michael E Goddard, Bonnie Berger, and Alkes L Price. Two variance component model improves genetic prediction in family data sets. *bioRxiv*, page 016618, 2015.

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

[13] 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.

[14] 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.

[15] 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.

[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] 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.

[18] Paul IW de Bakker, Manuel AR Ferreira, Xiaoming Jia, Benjamin M Neale, Soumya Raychaudhuri, and Benjamin F Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics*, 17(R2):R122–R128, 2008.

[19] 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.

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

[21] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, Douglas M Ruderfer, Andrew McQuillin, Derek W Morris, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

[22] 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.

[23] 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.

[24] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[25] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *arXiv preprint arXiv:1410.4803*, 2014.

Acknowledgements

Thanks B Neale, A Price, P Loh and B Vilhjalmsson and L Souchong for helpful comments.