Genomic Prediction and Genome Wide Association in Humans with Whole Genome Sequence Data

A. L. Price[1] **and P. Loh**[1]
[1]Harvard School of Public Health, Boston, USA

ABSTRACT: The transition from GWAS chip to sequencing data with increasingly larger sample sizes has many ramifications for efforts to conduct genomic prediction and genome wide association studies. First, as data sets grow larger, it is of interest to consider methods whose running time is linear in the data size. Second, it can be beneficial to model non-infinitesimal genetic architectures whose distribution of effect sizes is different from Gaussian, including minor allele frequency (MAF) dependent architectures. Third, although the fact that mixed model association can be viewed as a test for association on phenotypic residuals of BLUP predictions motivates a generalization to phenotypic residuals of predictions based on non-infinitesimal genetic architectures, this will require new approaches to calibration of test statistics. In this invited talk, we review recently published work in all of these research directions.
Keywords: prediction; genome wide association; sequence data; genetics

Introduction

Genomic prediction and genome wide association in humans have had substantial success in studies based on GWAS chip data (Visscher et al. (2012), Chatterjee et al. (2013)). However, as the costs of generating sequencing data decrease, GWAS chip studies will increasingly be replaced by whole-genome sequencing studies (The 1000 Genomes Project Consortium (2012), Pasaniuc et al. (2012)). Here, we highlight some of the issues related to this transition.

First, many widely used approaches for conducting genomic prediction and genome wide association using mixed model methods have running time $O(MN^2)$, where M is the number of markers and N is the number of samples (de los Campos et al. (2010), Yang et al. (2014)). However, as both M and N grow large, this running time may become computationally intractable, motivating the development of methods whose running time is only $O(MN)$, or linear in the data size.

Second, it can be beneficial to model non-infinitesimal genetic architectures whose distribution of effect sizes is different from Gaussian (Meuwissen et al. (2001), de los Campos et al. (2010)). Of particular interest to sequencing data, which includes a large number of rare variants, are genetic architectures in which the variance explained by a SNP depends on the MAF of that SNP (Speed et al. (2012), Lee et al. (2013)).

Third, standard mixed model association methods can be viewed as a test for association on phenotypic residuals of Best Linear Unbiased Predictions (BLUP) (Svishcheva et al. (2012)). This motivates the development of tests for association on phenotypic residuals of generalizations of BLUP predictions that model non-infinitesimal genetic architectures (Bolormaa et al. (2013)). However, if the goal is to produce calibrated test statistics that follow a specified null distribution, new approaches to calibration are needed, as the calibration approach of standard mixed model association methods does not generalize.

Materials and Methods

Genomic prediction. Let M be the number of markers and let N be the number of training samples. Let $\mathbf{X}$ be an M x N matrix of training sample genotypes, normalized to mean 0 and variance 1 for each SNP. The genetic relationship matrix (GRM) of training samples is defined as $\mathbf{A} = \mathbf{X}^T\mathbf{X}/M$. (The GRM can also be defined using pedigree relationships in data sets of related samples with known pedigrees (Henderson (1975)), but here we only consider the GRM defined from genetic data.) The heritability explained by genotyped SNPs (h_g^2) can be defined in the population as the maximum proportion of phenotypic variance that can be explained by a linear combination of genotyped SNPs, and can be estimated from the training samples by using restricted maximum likelihood (REML) to fit the phenotypic covariance matrix $\mathbf{V} = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}$ to the observed phenotypes $\mathbf{Y}$ and setting $h_g^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ (Yang et al. (2010)). The BLUP prediction of genetic values of training samples is $\mathbf{Y}_{BLUP} = h_g^2\mathbf{A}\mathbf{V}^{-1}\mathbf{Y}$ (or equivalently $\mathbf{Y}_{BLUP} = h_g^2(\mathbf{X}^T\mathbf{X}/M)\mathbf{V}^{-1}\mathbf{Y}$), and the BLUP prediction for a set of validation samples is $\mathbf{Y}^*_{BLUP} = h_g^2(\mathbf{X}^{*T}\mathbf{X}/M)\mathbf{V}^{-1}\mathbf{Y}$, where $\mathbf{X}^*$ is an M x N^* matrix of normalized genotypes in N^* validation samples (de los Campos et al. (2010)).

Mixed model association. Let $\mathbf{V} = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}$ and $\mathbf{Y}$ denote observed phenotypes as above. Let $\mathbf{X}_m$ denote an M x 1 vector of normalized genotypes at candidate SNP m. The standard mixed model association test is to compute a χ^2 statistic $t^2 = (\mathbf{X}_m^T\mathbf{V}^{-1}\mathbf{Y})^2/(\mathbf{X}_m^T\mathbf{V}^{-1}\mathbf{X}_m)$ (Yang et al. (2014)). The numerator of t^2 is proportional to $(\mathbf{X}_m^T(\mathbf{Y} - \mathbf{Y}_{BLUP}))^2$, since $\mathbf{Y} - \mathbf{Y}_{BLUP} = \mathbf{Y} - h_g^2\mathbf{A}\mathbf{V}^{-1}\mathbf{Y} = \mathbf{Y} - (\mathbf{V} - (1 - h_g^2)\mathbf{I})\mathbf{V}^{-1}\mathbf{Y} = (1 - h_g^2)\mathbf{V}^{-1}\mathbf{Y}$, and the denominator of t^2 is approximately constant (Svishcheva et al. (2012)). Thus, the mixed model association test can be viewed as an association test on the BLUP residual $\mathbf{Y} - \mathbf{Y}_{BLUP}$, although this is contingent on appropriate calibration (Svishcheva et al. (2012)).

Results and Discussion

Running Time. Direct solution of the mixed model equations requires $O(MN^2)$ time to compute the GRM and $O(N^3)$ time to invert the phenotypic covariance matrix. However, iterative methods circumvent the need to directly solve the mixed model equations. Iterative methods have been applied to pedigree-based mixed model analysis since the 1980s (Schaeffer and Kennedy (1986), Misztal and Gianola (1987), Berger et al. (1989)) and have recently

been extended to SNP-based mixed model analysis, yielding O(*MN*) time algorithms for BLUP prediction (Legarra and Misztal (2008), VanRaden (2008)) (Table 1). These methods apply various iterative linear algebra approaches (e.g., Gauss-Seidel iteration, Jacobi iteration, or preconditioned conjugate gradients) requiring only an O(*MN*) time matrix-vector multiplication at each iteration. Convergence typically occurs within a few dozen iterations.

Table 1. Published methods for genomic prediction with running time O(*MN*).

Reference	Approach	Effect size prior
Legarra and Misztal (2008)	GS	Normal
VanRaden (2008)	J	Normal
Meuwissen et al. (2009)	VB / ICE	Zero-exponential mix
Logsdon et al. (2010)	VB / ICE	Zero-trunc. normal mix
Carbonetto and Stephens (2012)	VB / ICE	Zero-normal mix
Logsdon et al. (2012)	VB / ICE	Zero-improper mix

GS: Gauss-Seidel iteration. J: Jacobi iteration. VB: Variational Bayes. ICE: Iterated conditional expectation.

Non-infinitesimal genetic architectures. BLUP methods are based on an infinitesimal (Gaussian) genetic architecture, but Meuwissen et al. (2001) proposed a Bayesian approach that assigns a non-infinitesimal prior distribution to additive SNP effects and obtains posterior estimates using Markov chain Monte Carlo (MCMC). In the past decade, numerous extensions of this approach have been developed (Erbe et al. (2012), Zhou et al. (2013), Gianola (2013)). In addition, recent progress has included fast methods that obtain approximate posterior estimates in O(*MN*) time (Table 1). These methods modify the update step used in the O(*MN*) time Gauss-Seidel infinitesimal mixed model approach (Legarra and Misztal (2008)) by iteratively updating each SNP effect with its conditional posterior mean, an approach that is variously described as "iterated conditional expectation" (Meuwissen et al. (2009)) or "variational Bayes" (Logsdon et al. (2010), Carbonetto and Stephens (2012), Logsdon et al. (2012)). The methods use similar computational approaches but differ in their assumed priors, methods of estimating hyperparameters, and approaches to model selection or averaging.

The shift from GWAS chip to sequencing data has the potentially to substantially improve prediction accuracy, due to rare causal variants not tagged by common SNPs (Meuwissen and Goddard (2010), Yang et al. (2010)). However, both the magnitude of the available improvement and the methods for capturing this improvement will depend on the relationship between the variance explained by a SNP and the MAF of that SNP (Speed et al. (2012), Lee et al. (2013)), which is a function of the strength of negative selection (against new mutations) on the trait (Agarwala et al. (2013)). Strong negative selection is required in order for the variance explained to be independent of the MAF p (a common model assumption), but in the absence of selection the variance explained will be proportional to $p(1-p)$. In the latter case, rare variants may not explain the gap between the heritability explained by genotyped SNPs in GWAS chip data (h_g^2) and the total narrow-sense heritability (h^2), and other explanations may be required (Zuk et al. (2012), Zaitlen et al. (2013)).

Generalizing mixed model association. The standard mixed model association test can be viewed as an association test on the BLUP residual $\mathbf{Y} - \mathbf{Y}_{\mathrm{BLUP}}$, as described above. The development of Bayesian methods that model non-infinitesimal genetic architectures to compute a prediction with increased accuracy ($\mathbf{Y}_{\mathrm{BAYES}}$) raises the question of whether association tests based on the residual $\mathbf{Y} - \mathbf{Y}_{\mathrm{BAYES}}$ may achieve higher power. Indeed, Bolormaa et al. (2013) showed that the resulting effect sizes are likely to be more precise, although they did not explore the question of calibration of test statistics. Logsdon et al. (2012) proposed a test statistic heuristically calibrated to the data, analogous to genomic control (Devlin and Roeder (1999)), but recent work has shown that genomic control is not an appropriate form of calibration at large sample sizes because test statistics are expected to be inflated by true polygenic signal (Yang et al. (2011), Yang et al. (2014)). Thus, the question of calibration of a test statistic based on $\mathbf{Y} - \mathbf{Y}_{\mathrm{BAYES}}$ is currently unresolved. One promising direction of research involves a method of calibration that takes advantage of the relationship between test statistics and linkage disequilibrium (Bulik-Sullivan et al. (2014)), but the question of whether that approach will provide an appropriate calibration to mixed model association statistics and their extensions remains an open question.

Conclusion

In this review, we have highlighted the appeal of O(*MN*) methods for BLUP and its extensions, while raising two unanswered questions. The first question involves the relationship between variance explained and MAF, which will vary across traits and can only be resolved empirically. The second question involves calibration of mixed model association statistics and their extensions, which is an important avenue for future work.

Acknowledgments

We are grateful to B. Vilhjalmsson, N. Patterson, B. Bulik-Sullivan, H. Finucane and B. Neale for helpful discussions. This work was funded by NIH grants R01 HG006399 and R01 MH101244.

Literature Cited

Agarwala, V., Flannick, J., Sunyaev, S. et al. (2013). Nat Genet. 45:1418-1427.

Berger, P., Luecke, G., and Hoekstra, J. (1989). J. Dairy Sci. 72:514-522.

Bolormaa, S., Pryce, J. E., Kemper, K. et al. (2013). J. Anim. Sci. 91:3088-3104.

Bulik-Sullivan, B. K., Loh, P., Finucane, H. et al. (2014). http://biorxiv.org/content/early/2014/02/21/002931. Accessed on Feb. 26, 2014.

Carbonetto, P., and Stephens, M. (2012). Bayesian Anal. 7(1):73-108.

Chatterjee, N., Wheller, B., Sampson, J. et al. (2013). Nat. Genet. 45:400-405.

de los Campos, G., Gianola, D., and Allison, D. B. (2010). Nat. Rev. Genet. 11:880-886.
Devlin, B., and Roeder, K. (1999). Biometrics 55:997-1004.
Erbe, M., Hayes, B. J., Matukumalli, L. K. et al. (2012). J. Dairy Sci. 95:4114-4129.
Gianola, D. (2013). Genetics 194:573-596.
Henderson, C. R. (1975). Biometrics 31:423-447.
Lee, S. H., Yang, J., Chen., G. B. et al. (2013). Am. J. Hum. Genet. 93:1151-1155.
Legarra, A., and Misztal, I. (2008). J. Dairy Sci. 91:360-366.
Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). BMC Bioinfo. 11:58.
Logsdon, B. A., Carty, C. L., Reiner, A. P. et al. (2012). Bioinformatics 38:1738-1744.
Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Genetics 157:1819-1829.
Meuwissen, T. H. E., Solberg, T. R., Shepherd, R. et al. (2009). Genet. Sel. Evol. 41(2):1-10.
Meuwissen, T., and Goddard, M. (2010). Genetics 185:623-631.
Misztal, I., and Gianola, D. (1987). J. Dairy Sci. 70:716-723.
Pasaniuc, B., Rohland, N., McLaren, P. J. et al. (2012). Nat. Genet. 44:631-635.
Schaeffer, L. R., and Kennedy, B. W. (1986). J. Dairy Sci. 69:575-579.
Speed, D., Hemani, G., Johnson, M. R. et al. (2012). Am. J. Hum. Genet. 91:1011-1021.
Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M. et al. (2012). Nat. Genet. 44:1166-1170.
The 1000 Genomes Project Consortium (2012). Nature 491:56-65.
VanRaden, P. M. (2008). J. Dairy Sci. 91:4414-4423.
Visscher, P. M., Brown, M. A., McCarthy, M. I. et al. (2012). Am. J. Hum. Genet. 90:7-24.
Yang, J., Benyamin, B., McEvoy, B. P. et al. (2010). Nat. Genet. 42:565-569.
Yang, J., Weedon, M. N., Purcell, S. et al. (2011) Eur. J. Hum. Genet. 19:807-812.
Yang, J., Zaitlen, N. A., Goddard, M. E. et al. (2014). Nat. Genet. 46:100-106.
Zaitlen, N., Kraft, P., Patterson, N. et al. (2013). PLoS Genet. 9:e1003993.
Zhou, X., Carbonetto, P., and Stephens, M. (2013). PLoS Genet. 9:e1003264.
Zuk, O., Hechter, E., Sunyaev, S. R. et al. (2012). Proc. Natl. Acad. Sci. U. S. A. 109:1193-1198.