LINK TO ORIGINAL ARTICLE
LINK TO INITIAL CORRESPONDENCE

Author reply to *A commentary on* Pitfalls of predicting complex traits from SNPs

Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard and Peter M. Visscher

Following our recent Opinion article (Pitfalls of predicting complex traits from SNPs. *Nature Rev. Genet.* **14**, 507–515 (2013))[1], we received correspondence by de los Campos and Sorensen (A commentary on *Pitfalls of predicting complex traits from SNPs. Nature Rev. Genet.* **14**, 894 (2013))[2]. We thank them for their comments, which follows their recent work[3]. de los Campos and Sorensen agree that maximum prediction accuracy depends on h^2_M, which is defined as the variance explained by genotyped markers in the population. They claim that estimates of h^2_M in a finite sample ($h^2_{G\text{-}BLUP}$ or h^2_G) may overestimate h^2_M, and that this is exacerbated for unrelated individuals. We respond by showing how and why we disagree with these claims.

h^2_G and $h^2_{G\text{-}BLUP}$ are estimates of the same parameter from equivalent models[4–7] and so, for the same data set, they must have the same value. Both measure the proportion of the phenotypic variance that is explained by the markers. This proportion depends on linkage disequilibrium (LD) between the single-nucleotide polymorphisms (SNPs) and causal variants (also known as quantitative trait loci (QTLs)). If the LD is imperfect, then h^2_M will be less than the conventional heritability (h^2), which is the proportion of variance explained by all causal variants. The extent of LD depends on the relatedness of the sample of individuals used. If closely related individuals are included in the sample, there is long-range LD generated even between SNPs and QTLs on different chromosomes. Thus, inclusion of close relatives increases h^2_M and its estimates. Usually, the parameter we wish to estimate is the h^2_M among individuals who are no more closely related than randomly sampled individuals from the population[8].

de los Campos and Sorensen state that the accuracy of prediction (R^2_{TST}) does not approach h^2_M even in an infinite sample. This is incorrect. R^2_{TST} depends on two factors — h^2_M and the accuracy with which the marker effects are estimated[4,9]. If the marker effects are estimated with no error, then $R^2_{TST} = h^2_M$. In practice, the accuracy of estimating SNP effects is usually low in humans, and this also explains the low R^2_{TST} that is often reported. Their recent study[3] claims that "the estimated h^2_G did not provide a good indication of prediction R^2". In their simulations of unrelated individuals (GEN cohort; $h^2 = 0.8$), they state that "when [non-causal] markers were used we observed only a small extent of missing heritability [$h^2_G = 0.737$, versus $h^2_G = 0.773$ for causal markers] but the reduction in R^2 due to use of markers that were in imperfect LD with causal loci was dramatic [$R^2 = 0.071$, versus $R^2 = 0.517$ for causal markers]". Even though the number of causal loci was the same, the number of markers differed: 300,000, corresponding to M = 60,000 independent markers versus M = 5,000 in the causal set. The following equation[1] (where N_d is the sample size in the discovery sample) demonstrates that R^2 decreases with higher M (which increases the variance of the estimated genetic relationships).

$$R^2 = \frac{h^2_M}{1 + \frac{M}{N_d h^2_M}(1 - R^2)}$$

de los Campos and Sorensen say that R^2_{TST} is zero if the training and testing data sets are independent. This is a distracting statement because individuals within a species are always related to some degree. They also question our focus on the prediction accuracy that can be obtained in an independent validation sample. We disagree with the opinion of de los Campos and Sorensen that the prediction accuracy that can be obtained in a non-independent validation sample is a quantity of equal interest.

Naomi R. Wray, Jian Yang and Peter M. Visscher are at The Queensland Brain Institute, The University of Queensland, QBI Building, St Lucia, Queensland 4071, Australia.

Jian Yang and Peter M. Visscher are at The University of Queensland Diamantina Institute, Level 7, 37 Kent Street, Translational Research Institute, Woolloongabba, Queensland 4102, Australia.

Ben J. Hayes and Mike E. Goddard are at the Biosciences Research Division, Department of Primary Industries, GPO Box 4440, Melbourne, Victoria 3001, Australia.

Ben J. Hayes is at the Dairy Futures Cooperative Research Centre, AgriBio, Centre for AgriBioscience, 5 Ring Road, La Trobe University, Bundoora, Victoria 3083, Australia; and La Trobe University, Bundoora, Victoria 3086, Australia.

Alkes L. Price is at the Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA; the Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA; the Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; and the Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA.

Mike E. Goddard is at the Faculty of Land and Food Resources, University of Melbourne, Melbourne, Victoria 3010, Australia.

Correspondence to P.M.V.
e-mail: peter.visscher@uq.edu.au

doi:10.1038/nrg3457-c2
Published online 18 November 2013

1. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nature Rev. Genet.* **14**, 507–515 (2013).
2. de Los Campos, G. & Sorensen, D. A. A commentary on *Pitfalls of predicting complex traits from SNPs*. *Nature Rev. Genet.* **14**, 894 (2013).
3. de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C. & Sorensen, D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* **9**, e1003608 (2013).
4. Goddard, M. E., Wray, N. R., Verbyla, K. L. & Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Statist. Sci.* **24**, 517–529 (2009).
5. Habier, D., Fernando, R. L. & Dekkers, J. C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397 (2007).
6. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
7. Goddard, M. E. Genomic Selection: predicion of accuracy and maximisation of long term response. *Genetica* **136**, 245–257 (2009).
8. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
9. Goddard, M. E., Hayes, B. J. & Meuwissen, T. H. Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed.Genet.* **128**, 409–421 (2011).
10. Makowsky, R. *et al.* Beyond missing heritability: prediction of complex traits. *PLoS Genet.* **7**, e1002051 (2011).
11. Janss, L., de Los Campos, G., Sheehan, N. & Sorensen, D. Inferences from genomic models in stratified populations. *Genetics* **192**, 693–704 (2012).

Competing interests statement
The authors declare no competing interests.