Published in final edited form as:

Nat Rev Genet. 2013 December ; 14(12): 894. doi:10.1038/nrg3457-c2.

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

Naomi R. Wray¹, Jian Yang^{1,2}, Ben J. Hayes^{3,4}, Alkes L. Price⁵, Michael E. Goddard^{3,6}, and Peter M. Visscher^{1,2}

¹The Queensland Brain Institute, The University of Queensland, QBI Building, St Lucia, Queensland 4071, Australia

²The University of Queensland Diamantina Institute, Level 7, 37 Kent Street, Translational Research Institute, Woolloongabba, Queensland 4102, Australia

³Biosciences Research Division, Department of Primary Industries, GPO Box 4440, Melbourne, Victoria 3001, Australia

⁴Dairy 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

⁵Department 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

⁶University of Melbourne, Melbourne, Victoria 3010, Australia

We thank de los Campos and Sorensen (D&S) for their Correspondence, which follows their recent work¹. D&S agree that maximum prediction accuracy depends on h^2_M , 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-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-BLUP} are estimates of the same parameter from equivalent models^{3–6} and so, for the same dataset, 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 SNPs and causal variants or quantitative trait loci (QTL). 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 QTL 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 (e.g.⁷).

D&S 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^{3; 8}. 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} often reported. Ref.¹ 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 “when [non-causal] markers were used we observed only a small extent of missing heritability [$h^2_G=0.737$, vs. $h^2_G=0.773$ for causal markers] but the reduction in R^2 due to use of markers that were in imperfect LD with causal loci was dramatic [$R^2=0.071$, vs. $R^2=0.517$ 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 markers² vs $M=5,000$ in the causal set. Equation (1) of ref.² demonstrates that R^2 decreases with higher M (which increases the variance of the estimated genetic relationships).

D&S say that R^2_{TST} is zero if the training and testing datasets are independent. This is a distracting statement because individuals within a species are always related to some degree. D&S also question our focus on the prediction accuracy that can be obtained in an independent validation sample². We disagree with the opinion of D&S that the prediction accuracy that can be obtained in a non-independent validation sample is a quantity of equal interest.

References

1. de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 2013; 9:e1003608. [PubMed: 23874214]
2. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet.* 2013; 14:507–515. [PubMed: 23774735]
3. Goddard ME, Wray NR, Verbyla KL, Visscher PM. Estimating effects and making predictions from genome-wide marker data. *Statistical Science.* 2009; 24:517–529.
4. Habier D, Fernando RL, Dekkers JC. The impact of genetic relationship information on genome-assisted breeding values. *Genetics.* 2007; 177:2389–2397. [PubMed: 18073436]
5. VanRaden PM. Efficient methods to compute genomic predictions. *Journal of dairy science.* 2008; 91:4414–4423. [PubMed: 18946147]
6. Goddard ME. Genomic Selection: prediction of accuracy and maximisation of long term response. *Genetica.* 2009; 136:245–257. [PubMed: 18704696]
7. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010; 42:565–569. [PubMed: 20562875]
8. Goddard ME, Hayes BJ, Meuwissen TH. Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of animal breeding and genetics = Zeitschrift für Tierzucht und Zuchtungsbiologie.* 2011; 128:409–421. [PubMed: 22059574]

9. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de Los Campos G. Beyond missing heritability: prediction of complex traits. *PLoS Genet.* 2011; 7:e1002051. [PubMed: 21552331]
10. Janss L, de Los Campos G, Sheehan N, Sorensen D. Inferences from genomic models in stratified populations. *Genetics.* 2012; 192:693–704. [PubMed: 22813891]