LETTER

GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs

Jian Yang[a,b,1], S. Hong Lee[c], Naomi R. Wray[a], Michael E. Goddard[d,e], and Peter M. Visscher[a,b,1]

In a recent publication in PNAS, Krishna Kumar et al. (1) claim that "GCTA applied to current SNP data cannot produce reliable or stable estimates of heritability." We show below that those claims are false due to their misunderstanding of the theory and practice of random-effect models underlying genome-wide complex trait analysis (GCTA) (2).

GCTA, more precisely, the genomic-relatedness-based restricted maximum-likelihood (GREML) approach (3) implemented in GCTA (4), is a method to estimate the proportion of phenotypic variation that can be explained by all genome-wide SNPs (h^2_g) using an SNP-derived genetic relationship matrix. Krishna Kumar et al. (1) claim that the estimate of h^2_g from GCTA-GREML is unreliable based on the observations that the observed variance explained per SNP ($\sigma^2 = h^2_g/m$, where m is the number of SNPs) from simulations is inconsistent with their expectation. This is because they misunderstand that "GCTA assumes that the SNPs used are in linkage equilibrium" (ref. 1, p. 2), and mistakenly believe that σ^2 should be the same regardless of the number of SNPs fitted in the model in either their original paper (1) or subsequent response (5) to our commentary (2). In fact, GREML fits all of the SNPs jointly in a random-effect model so that each SNP effect is fitted conditioning on the joint effects of all of the other SNPs [i.e., it accounts for linkage disequilibrium (LD) between the SNPs] (3). The estimate of σ^2 in a random effect is interpreted as the variance of an SNP effect when it is fitted jointly with all of the other SNPs. Therefore, σ^2 for a random subset of SNPs (σ^2_{subset}) is larger than that for the entire set (σ^2_{entire}) if SNPs are in LD.

Krishna Kumar et al. (1) show by analysis of a real dataset (ref. 1, figure 4*A*) that $\hat{\sigma}^2_{subset}$ was, on average, much larger than $\hat{\sigma}^2_{entire}$. They further used the estimates from our previous studies (3, 6) as examples to show that $\hat{\sigma}^2$ with a smaller m was larger than with a larger m (5). All of these observations are entirely consistent with published theory that $\hat{\sigma}^2_{subset}$ for a random subset of SNPs is larger than $\hat{\sigma}^2_{entire}$ if SNPs are in LD. It was clearly demonstrated by Yang et al. (figure 2 of ref. 3) that h^2_g increases toward a plateau as m increases.

From simulations of unlinked SNPs (figure 2 in ref. 1), Krishna Kumar et al. (1) observed that SD($\hat{\sigma}^2_{subset}$) was much larger than SD($\hat{\sigma}^2_{entire}$). Their claim that this is a failure of GCTA-GREML is therefore incorrect because SD($\hat{\sigma}^2$) is expected to increase with a decrease in m. If SNPs are unlinked, $\mathrm{SD}(\hat{\sigma}^2) \approx \frac{1}{N}\sqrt{\frac{2}{m}}$, where n = sample size (2, 7). For n = 2,000 and m = 50,000, this equation predicts that $\mathrm{SD}(\hat{\sigma}^2) \approx 3.2 \times 10^{-6}$, which is highly consistent with the observation by Krishna Kumar et al. (1) of 3.1×10^{-6}.

There are many other errors in the paper by Krishna Kumar et al. (1), as pointed out by us (2) and others (8). In conclusion, Krishna Kumar et al. (1, 5) misunderstood the model and assumptions underlying GCTA-GREML, and therefore used the incorrect expected mean and SD of $\hat{\sigma}^2_{subset}$ for comparison with those values observed from resampling. Hence, their conclusion about biasedness of GREML estimates is not supported by empirical evidence.

Acknowledgments

We thank Bill Hill, Alkes Price, John Witte, and Mark Blows for comments and support. This research was supported by the Australian National Health and Medical Research Council (Grants 1078037 and 1078901) and the Sylvia & Charles Viertel Charitable Foundation.

1 Krishna Kumar S, Feldman MW, Rehkopf DH, Tuljapurkar S (2016) Limitations of GCTA as a solution to the missing heritability problem. *Proc Natl Acad Sci USA* 113(1):E61–E70.

2 Yang J, Lee SH, Wray NR, Goddard ME, Visscher PM (2016) Commentary on "Limitations of GCTA as a solution to the missing heritability problem." *bioRxiv*, 10.1101/036574.

[a]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia; [b]The University of Queensland Diamantina Institute, The Translation Research Institute, Brisbane, QLD 4102, Australia; [c]School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia; [d]Faculty of Veterinary and Agricultural Science, University of Melbourne, Parkville, VIC 3010, Australia; and [e]Biosciences Research Division, Department of Economic Development, Jobs, Transport and Resources of Bundoora, Bundoora, VIC 3083, Australia

Author contributions: J.Y. and P.M.V. designed research; S.H.L. and N.R.W. contributed new reagents/analytic tools; and J.Y., M.E.G., and P.M.V. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: jian.yang@uq.edu.au or peter.visscher@uq.edu.au.

3 Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569.
4 Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1):76–82.
5 Kumar SK, Feldman MW, Rehkopf DH, Tuljapurkar S (2016) Response to Commentary on "Limitations of GCTA as a solution to the missing heritability problem." *bioRxiv*, 10.1101/039594.
6 Yang J, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 43(6):519–525.
7 Visscher PM, et al. (2014) Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* 10(4):e1004269.
8 Gamazon ER, Park DS (2016) SNP-based heritability estimation: Measurement noise, population stratification and stability. *bioRxiv*, 10.1101/040055.