PhD rotation project 1

Evaluating evidence for non-additive effects within loci associated with inflammatory bowel disease risk

Benjamin Y. H. Bai*

Supervisors: Dr Carl Anderson and Dr Jeffrey Barrett

Genome-wide association (GWA) studies have been essential in unravelling the architecture of complex, common-variant diseases[1,2]. In inflammatory bowel disease (IBD) and its two major forms: Crohn's disease (CD) and ulcerative colitis (UC), at least 240 loci are implicated to date[3]. Although there is evidence of non-additive effects in IBD[4,5], all implicated loci were identified assuming an additive genetic model. Knowledge of the correct genetic model underlying associations—especially at lead variants—is important for fine-mapping[6] and meta-analyses[7]. We searched for non-additive effects within implicated loci in a well-studied 2016 UK IBD cohort[3]. Non-additivity was detected in the major histocompatibility complex (MHC) region, with most signal being UC-specific, corroborating previous findings[4]. We found little evidence for non-additivity at lead variants, consistent with the expectation that additive effects are most prominent in complex disease[8]. A more comprehensive analysis, incorporating other existing IBD cohorts[9,10] for additional power will be required to draw conclusions at a genome-wide level.

In association studies, the assumed genetic model defines the relationship between the number of copies, or dosage of a disease-associated allele and the resulting increase in risk. Without prior information, it is biologically-plausible to expect the relative risk of carrying one copy of a disease-associated allele to lie between the baseline and the risk conferred by carrying two copies. Most studies use a one degree-of-freedom (1 d.f.) additive model, where risk—measured as the log-odds of an individual being affected—increases linearly with dosage.

The additive model is generally robust to small departures from linearity, without the problem of multiple testing introduced by considering additional models in parallel[11]. Also, there is often little data available to estimate the effect of rare homozygotes, hence the additive model is recommended for initial screening[12]. All 240 loci reported in the latest IBD meta-analysis[3] were detected assuming additivity.

However, the additive model performs poorly when the mode of effect is heterozygotic[11], and for recessive effects with low minor allele frequency (MAF)[13]. Possible 1 d.f. non-additive models include dominant, recessive, and heterozygotic[14]. Power to detect association is greatest if the correct underlying model is assumed[13]. When the mode of effect is unknown, one can fit a 2 d.f. general model, where additional copies of the disease-associated allele confer risk independently[14]. The need to estimate an extra parameter over the 1 d.f. model results in less powerful tests at the majority of loci which have additive or near-additive effects, thus the 2 d.f. test is not in widespread usage. One can often achieve better power by considering several 1 d.f. models, even accounting for multiple testing[13,15].

*Wellcome Trust Sanger Institute, Hinxton, UK. Oct-Dec 2016. *Contact:* bb9@sanger.ac.uk.

The recent emergence of large, well-studied cohorts[3,9] allows us to survey for non-additivity in IBD, where GWA studies are powered to detect variants down to 1% MAF[16]. There is some evidence for non-additivity in IBD, particularly in the MHC region for variants implicated in UC[4]. Also, previous linkage analyses suggest a recessive inheritance model at *NOD2*, a locus associated primarily with CD[5].

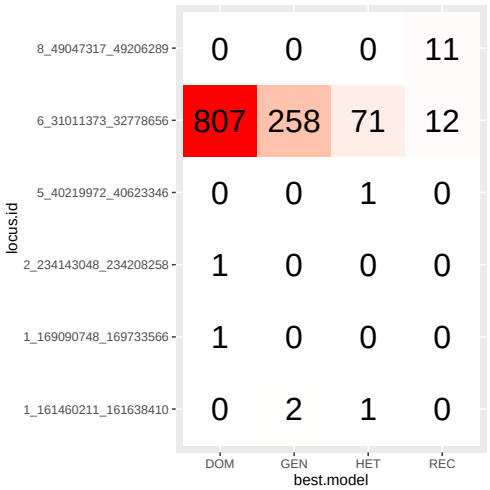
Here, we evaluate evidence for non-additive effects in a 2016 UK IBD cohort (GWAS3)[3], consisting of 9495 controls and 8860 cases (4264 CD, 4072 UC, 524 unclassified IBD) after quality control, typed over 296 thousand markers with MAF $> 0.1\%$. We conducted our analysis within the 240 known loci for which there is already genome-wide significant evidence for association in IBD, UC, or CD[3]—where we should be well-powered to detect deviations from additivity. Determining the correct mode of effect at these loci is important for avoiding misleading results from meta-analysis procedures that make assumptions about the underlying genetic model[7]. In particular, we focused on detecting non-additivity at lead variants (those with the smallest association p-value within each locus), which are central to fine-mapping techniques based on building credible sets of potentially-causal variants[6].

Analyses were performed over the entire IBD cohort, and for the CD-specific and UC-specific sub-cohorts. Frequentist association tests were run using SNPTEST[17] over all five available genetic models (additive, dominant, recessive, general, and heterozygotic). As the resulting association p-values are calculated given a specific model, they should not be used to distinguish between models. To facilitate model comparison, we replicated the association testing for the 163 thousand filtered sites (INFO > 0.4, MAF $> 1\%$) that fell within known loci using logistic regression. P-values from logistic regression and SNPTEST were highly concordant ($r > 0.95$ in all analyses).

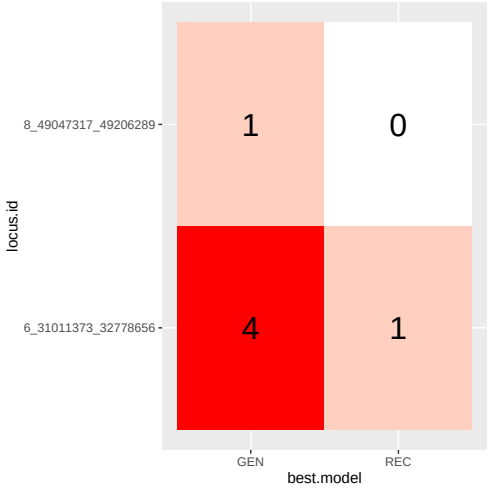
At each site, we determined the best model using the likelihood ratio test (LRT), Bayesian information criterion (BIC) and Vuong's non-nested hypothesis test[18,19] (see Methods for details). The distribution of non-additive best models for each variant that achieved genome-wide significance ($p < 5 \times 10^{-8}$) in at least one model is shown in Figure 1. Nearly all signal originated from a single locus on chromosome 6, and this signal was enriched in UC compared to CD. The locus is located with the MHC, a region previously identified to contain protective dominant, risk recessive, and overdominant effects, also primarily in UC[4].

We then searched for variants that were not lead variants when ranked by their additive p-values, but became lead variants when each variant was ranked by the BIC of its best model. We found apparent evidence for this at three sites, one of which was a known meta-analysis lead variant (Table 1). These sites had very significant p-values and large BIC decreases under their best models compared to the additive model. However, on inspecting their genotype cluster plots, we discovered these signals were false positives driven by incorrect genotype calls, resulting in extreme deviation from Hardy–Weinberg equilibrium in cases (Fig. 2). Such sites were not excluded by quality control procedures, as the prevailing practice is to filter for deviations from Hardy–Weinberg equilibrium in controls only, to avoid the possibility of excluding associated variants under selection[20]. These results suggest it may be prudent to introduce a very stringent filter in cases, although manual examination of cluster plots remains the best way to ensure calls are robust.

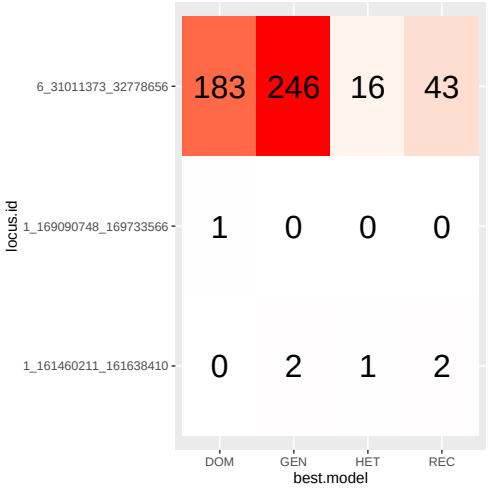
In our analysis of lead variants, non-additive variants in the MHC (from Figure 1) did not appear. Although the magnitude of their rank shifts were large, they had low ranks overall. For instance, the mean rank of the 488 MHC signals in UC (Fig. 1c) based on their additive p-values was 8625.10 (out of 33279 significant variants in the locus), which improved to 5157.46 based on the BIC of their best models, representing a mean BIC decrease of 13.16. This suggests non-additive effects within the region may be 'masked'



(a) IBD: 6567 significant variants within known loci (1165 with a non-additive best model).

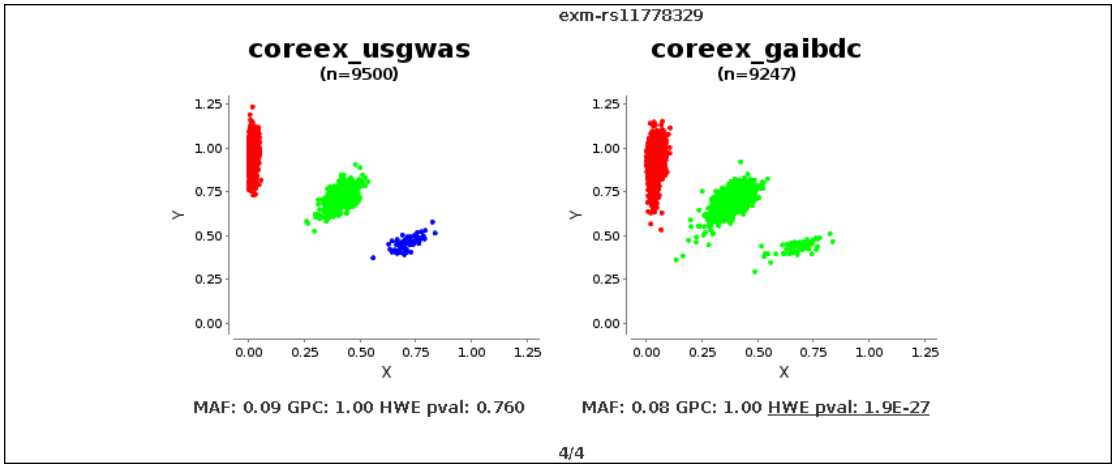


(b) CD: 1833 significant variants within known loci (6 with a non-additive best model).

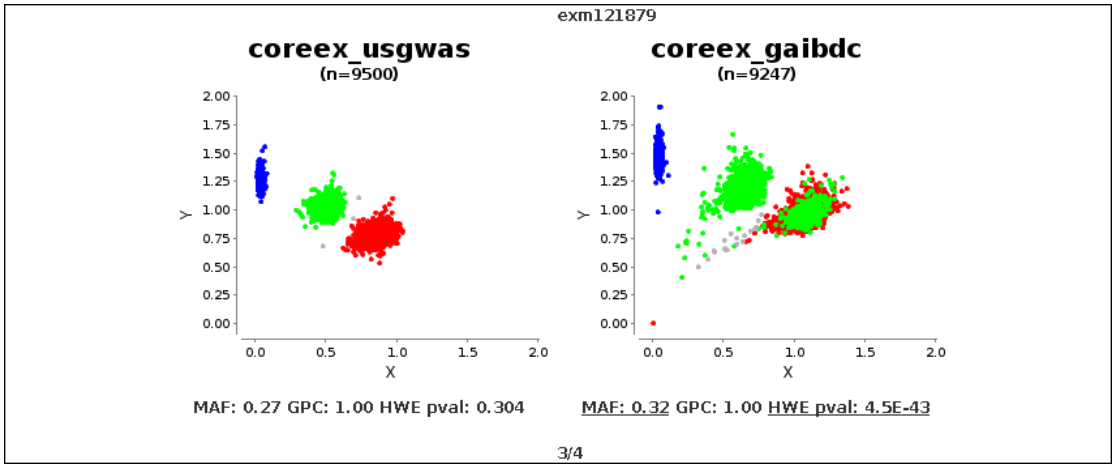


(c) UC: 8676 significant variants within known loci (494 with a non-additive best model).

Figure 1: Distribution of the best non-additive model for genome-wide significant variants within known IBD loci, by analysis (IBD, UC, CD), model (DOMinant, RECessive, HETerozygotic, GENeral) and locus (chr_leftCoord_rightCoord). Non-additivity is observed at a higher frequency in UC compared to CD. Most of the signals detected in CD were subsequently confirmed to be false-positives by inspection of their genotype cluster plots.



(a) Variant ID: 1:169511555_T_C



(b) Variant ID: 8:49053165_C_T

Figure 2: Genotype cluster plots for two of the variants in Table 1, separated into controls (coreex_usgwas, left) and cases (coreex_gaibdc, right). The three distinct clusters of points on each plot correspond to samples with each of the three possible genotypes at that site. Figure 2a shows a failure of the calling algorithm to distinguish between heterozygotes and minor allele homozygotes in cases. Figure 2b shows a batch effect that occured when a subset of cases were poorly called, then merged with correctly called batches downstream. The controls at both these sites demonstrate correct calling behaviour. MAF, genotyping rate (GPC) and the p-value for deviation from Hardy–Weinberg equilibrium are shown underneath each plot. Samples numbers are pre-quality control. Plots were generated using Evoker (v2.3)[21].

Table 1: Lead variants within known loci that arose only when variants were ranked by the BIC of their best models. The model mode of effect (DOMinant, RECessive, HETerozygotic, GENeral) and analysis in which the effect was discovered (IBD, UC, CD) are shown below. All three variants displayed a large reduction in p-value and BIC compared to the ADDitive model, and were subsequently confirmed to be false-positives by inspection of their genotype cluster plots (Fig. 2). Note that the best model p-value shown below is not necessarily the smallest p-value reported by SNPTEST e.g. for 8:49053165_C_T, the smallest p-value in CD was 3.80×10^{-08} (REC), but there was insufficient evidence that the recessive model fit better than the additive model.

Variant ID	Best model	Known?	ADD p-value	ADD rank	Best model p-value	BIC decrease
1:161479745_A_G	HET (IBD)	Yes	7.06×10^{-2}	177	1.59×10^{-14}	55.74
	GEN (UC)	Yes	1.88×10^{-2}	140	2.84×10^{-19}	70.49
1:169511555_T_C	DOM (IBD)	No	5.61×10^{-7}	2	1.09×10^{-9}	12.11
8:49053165_C_T	REC (IBD)	No	1.84×10^{-2}	12	1.75×10^{-14}	71.73
	GEN (CD)	No	5.26×10^{-2}	24	2.70×10^{-7}	32.78

by stronger, additive signals that occupy the higher ranks. It is known that there are multiple independent effects within the MHC[4]. Analyses conditioning on each of these effects in turn will be necessary to determine whether the non-additivity we observed is independent of these known signals[6].

Overall, we find that additive effects remain the most relevant effect type for common variants within known IBD loci, with the exception of the MHC region. However, only 30 of the 159 known IBD-specific loci, 26 of the 51 CD-specific loci, and 21 of the 35 UC-specific loci reached genome-wide significance in our analyses, thus we cannot discount the existence of non-additive effects outside these loci. A more powerful survey including other existing IBD cohorts[9,10] may be required, especially for drawing conclusions on a genome-wide context. Gathering enough data to non-additive effects at rare variants which are not well-represented in conventional array-based GWA studies remains a challenging proposition, although rare variants appear to play a relatively minor role in IBD[9].

Methods

Cohort samples were quality-controlled, phased, and imputed as detailed previously[3]. Association testing was performed using SNPTEST (v2.5)[17], conditioning on the first ten standardised sample principle components. All five available genetic models (ADD, DOM, REC, HET, GEN) were run. After association testing, we filtered out variants with INFO < 0.4 and MAF $< 1\%$, as there is low power for distinguishing between models at these sites.

To evaluate model fits, for each of the 163 thousand variants that lay within the 240 known loci[3], we fit a logistic regression using *glm* (R v3.30)—also conditioned on the first ten standardised sample principle components—where the dependent variable was the case-control phenotype, and the independent variable was a model-dependent allelic dosage term. This dosage term was calculated as $\mathbf{w} \cdot [g_{AA}, g_{Aa}, g_{aa}]$, where g_{AA}, g_{Aa}, and g_{aa} are the imputed genotype probabilities for the sample being homozygous reference, heterozygous, and homozygous alternate at that site respectively; and the weights $\mathbf{w}$ depend on the assumed model: additive $[0, 1, 2]$, dominant $[0, 1, 1]$, recessive $[0, 0, 1]$, and heterozygotic $[0, 1, 0]$. The general model was encoded as the additive model with the recessive dosage term included as a second independent variable. Association p-values were derived from the LRT against the null model with $[0, 0, 0]$ weights.

As the additive and general models are nested in our encoding, we also used the LRT to determine whether the general model was a significantly better fit. To distinguish between non-nested models (additive versus the other 1 d.f. models), we selected the inter-

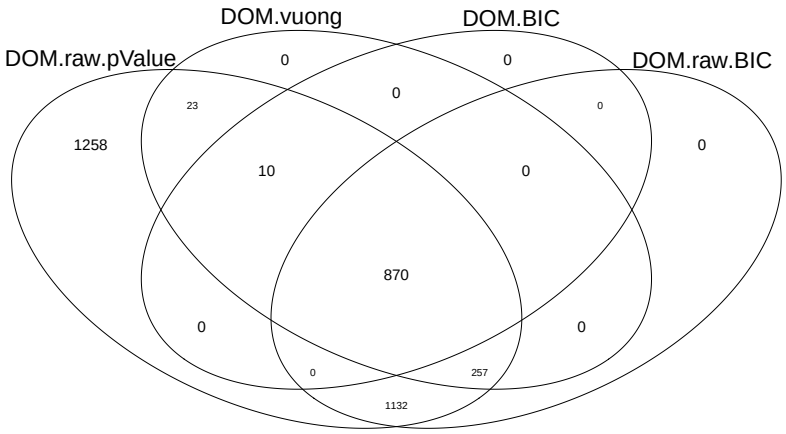
section of positive results based on the 95% confidence interval for the difference in BICs (a significant decrease in BIC being indicative of better fit[22]), and Vuong's non-nested hypothesis test (rejection of the null hypothesis where the additive and non-additive models fit equally well, using a Bonferroni-corrected α of $0.05/240$ loci)[18,19]. The intersection provided a more conservative result compared to selecting based on association p-value or raw BIC decrease (Fig. 3). The best model for a site was defined to be the model with the lowest BIC out of all models for which there was evidence of a better fit than the additive model, defaulting to the additive model if no such alternative was found.

Acknowledgements

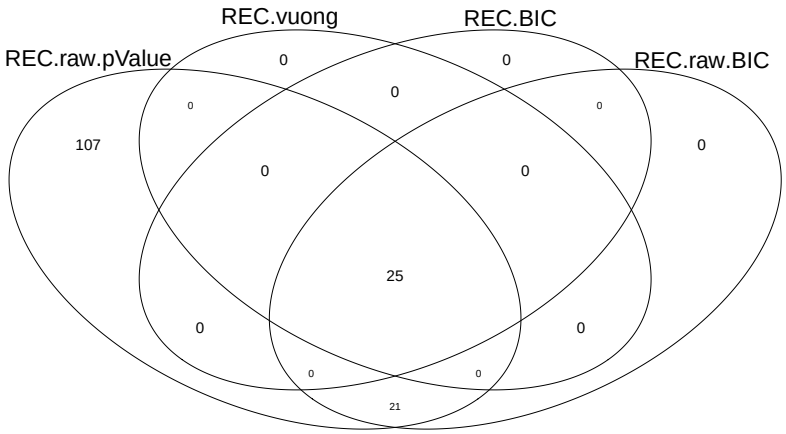
I thank my supervisors: Carl Anderson and Jeffrey Barrett; and the other members of the Anderson and Barrett teams at the Wellcome Trust Sanger Institute—in particular, Loukas Moutsianas, Katrina de Lange, Dan Rice, and John Lees—for their help and support during this rotation project.

References

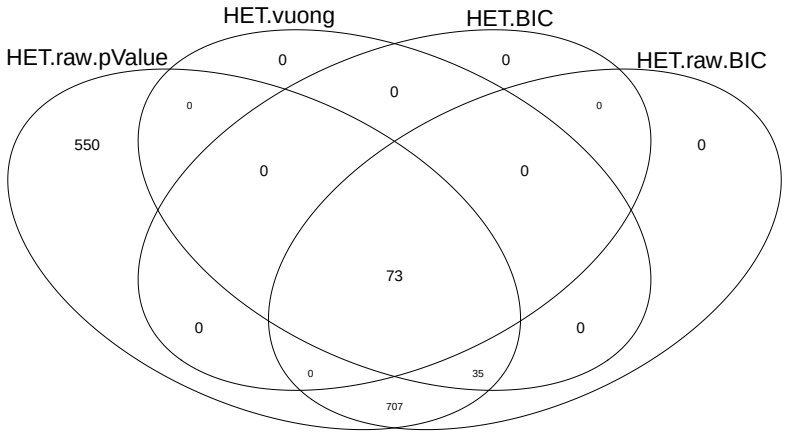
1. The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007).
2. McCarthy, M. I. M. *et al.:* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356–69 (2008).
3. De Lange, K. M. *et al.:* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *bioRxiv,* 1–19 (2016).
4. Goyette, P. *et al.:* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nature Genetics* **47**, 172–9 (2015).
5. Hugot, J. P. *et al.:* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
6. Spain, S. L. & Barrett, J. C.: Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24**, R111–R119 (2015).
7. Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J.: The choice of a genetic model in the meta-analysis of molecular association studies. *International Journal of Epidemiology* **34**, 1319–1328 (2005).
8. Hill, W. G., Goddard, M. E. & Visscher, P. M.: Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics* **4** (2008).
9. Luo, Y. *et al.:* Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at ADCY7. *bioRxiv* (2016).
10. Huang, H. *et al.:* Association mapping of inflammatory bowel disease loci to single variant resolution. *bioRxiv* (2015).
11. Dorak, M. T.: *Genetic Association Studies: Background, Conduct, Analysis, Interpretation* (2016).
12. Cantor, R. M., Lange, K. & Sinsheimer, J. S.: Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *American Journal of Human Genetics* **86**, 6–22 (2010).
13. Lettre, G., Lange, C., Hirschhorn, J. N. & Nicolae, D. L.: Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology* **31**, 358–362 (2007).
14. Tsepilov, Y. A. *et al.:* Nonadditive effects of genes in human metabolomics. *Genetics* **200**, 707–718 (2015).
15. González, J. R. *et al.:* Maximizing association statistics over genetic models. *Genetic Epidemiology* **32**, 246–254 (2008).
16. Khor, B., Gardet, A. & Xavier, R. J.: Genetics and pathogenesis of inflammatory bowel disease. eng. *Nature* **474**, 307–317 (2011).
17. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P.: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–13 (2007).



(a) Significant variants where the dominant model fits better.



(b) Significant variants where the recessive model fits better.



(c) Significant variants where the heterozygotic model fits better.

Figure 3: Distribution of the significant variants where, in IBD, a non-additive 1 d.f. model was deemed fit better than the additive model, based on four different model comparison methods: a decrease in SNPTEST p-value (raw.pValue), rejection of the null hypothesis that model fits are equal by Vuong's non-nested hypothesis test (vuong), a decrease in BIC with a 95% confidence interval that does not contain zero (BIC), and a decrease in BIC with magnitude greater than six (considered "strong" evidence[22]) (raw.BIC). The methods based on p-value and raw BIC decrease were avoided due to the large numbers of positives specific to those methods.

18. Vuong, Q. H.: Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society,* 307–333 (1989).

19. Merkle, E. C., You, D. & Preacher, K. J.: Testing nonnested structural equation models. *Psychological Methods* **20**, 1–22 (2015).

20. Anderson, C. A. *et al.:* Data quality control in genetic case-control association studies. *Nature Protocols* **5**, 1564–73 (2010).

21. Morris, J. A., Randall, J. C., Maller, J. B. & Barrett, J. C.: Evoker: A visualization tool for genotype intensity data. *Bioinformatics* **26**, 1786–1787 (2010).

22. Raftery, A. E.: Bayesian model selection in social research. *Sociological Methodology,* 111–163 (1995).