A note on simulating null distributions for G matrix comparisons

Michael B. Morrissey[1], *Sandra Hangartner*[2*] *and Keyne Monro*[2,3*]

12 July, 2019

Contact:

[1] corresponding author:

michael.morrissey@st-andrews.ac.uk

School of Biology

University of St Andrews

[2] School of Biological Sciences

Monash University

[3] Centre for Geometric Biology

Monash University

[*] these authors contributed equally

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/evo.13842

Abstract

Genetic variances and covariances, summarised in **G** matrices, are key determinants of the course of adaptive evolution. Consequently, understanding how **G** matrices vary among populations is critical to answering a variety of questions in evolutionary biology. A method has recently been proposed for generating null distributions of statistics pertaining to differences in **G** matrices among populations. The general approach facilitated by this method is likely to prove to be very important in studies of the evolution of **G**. We have identified an issue in the method that will cause it to create null distributions of differences in **G** matrices that are likely to be far too narrow. The issue arises from the fact that the method as currently used generates null distributions of statistics pertaining to differences in **G** matrices across populations by simulating breeding value vectors based on **G** matrices estimated from data, randomising these vectors across populations, and then calculating null values of statistics from **G** matrices that are calculated directly from the variances and covariances among randomised vectors. This calculation treats breeding values as quantities that are directly measurable, instead of predicted from **G** matrices that are themselves estimated from patterns of covariance among kin. The existing method thus neglects a major source of uncertainty in **G** matrices, which renders it anticonservative. We first suggest a correction to the method. We then apply the original and modified methods to a very simple instructive scenario. Finally, we demonstrate the use of both methods in the analysis of a real data set.

Key words: **G** matrix, quantitiatve genetics, null distribution, tensor analysis, differentiation

Introduction

Genetic variances and covariances among traits, summarised in **G** matrices, are central to understanding how phenotypes evolve (Walsh & Lynch 2018). Consequently, understanding similarities and differences in **G** among populations is critical for understanding whether, how, and by what rates and patterns, populations will diverge in response to prevailing patterns of natural selection. **G** is inherently challenging to estimate, and its multi-dimensional nature means that summarising differences in **G** matrices among populations, when each **G** matrix is typically imprecisely estimated, is a delicate and challenging task. Aguirre *et al.* (2014) propose a method for generating null distributions of statistics describing differences in **G** among populations, providing an important and unified advance in the available tools for studying **G** and its role in adaptive evolutionary diversification.

We have identified a feature of the Aguirre *et al.* (2014) method that is likely to render it anticonservative. In other words, the method as proposed, and currently applied, will generate null distributions of differences among **G** matrices that are likely to be too narrow, potentially inflating the apparent statistical significance of estimated differences among **G** matrices. First, we explain the cause of anticonservatism in the current method. We then suggest a modification to Aguirre *et al.*'s (2014) method to correct the issue. Next, we illustrate the anticonservatism of the original approach and the more desirable behaviour of the modified method with a simple example, wherein the desired properties of the null distribution are immediately apparent. Finally, we apply both methods to a more complex problem involving real data.

The original method and its modification

The method proposed by Aguirre *et al.* (2014) requires that (1) posterior distributions of estimated **G** matrices are generated for multiple populations, and (2) desired statistics summarising differences among the estimated **G** matrices are calculated. Null distributions of these statistics are then obtained by the following steps: (3) for each posterior sample in each population, breeding values for all individuals in the analysis are simulated by sampling from multivariate normal distributions, with covariances defined by the pedigree and population-specific estimates of **G**. Next, (4) all of the breeding values are re-assigned at random to all individuals, across populations. Then, (5) the covariances matrices of the randomised breeding values of individuals within each population are calculated. Finally (6) the statistics in step (2) are generated for each posterior sample from the randomised **G** matrices. This set of statistics serves as a null distribution, to which the actual values generated in step (2) can be compared.

The problem with this method arises in step (5), because there are two sources of statistical error in the

estimation of **G** matrices. The first source arises from the fact that in any finite study, the unknown additive genetic values and observed phenotypic values (even if the phenotypes are measured without error) of sampled individuals can imperfectly reflect the distributions of these values in the wider population from which individuals are drawn. This is the standard source of statistical error with which most statistical procedures are designed to grapple. Under ideal conditions (e.g., no measurement error or study-specific confounding factors), it is the only source of error in population-level summaries of quantities, such as phenotypes (or the phenotypic covariance matrix, **P**), that can be measured directly on individuals. Step (5) of Aguirre *et al.*'s (2014) method accounts for this source of variation only.

The second source of statistical error associated with **G** matrices arises from the fact that **G** is a population-level summary of additive genetic values that cannot be measured directly on individuals, despite being individual-level quantities. Rather, we must infer such values from patterns of similarity among kin. Consequently, **G** matrices are typically estimated less precisely than they would be if we could simply calculate variances and covariances among directly-measureable breeding values. This second source will typically be a major, if not the dominant, source of statistical error in estimates of **G**, and so a modification of the Aguirre *et al.* (2014) method is likely to be necessary in order for it to fully represent uncertainty in comparisons of **G** matrices.

We propose that the Aguirre *et al.* (2014) method can be improved by using steps (1) through (4) as described above, then replacing step (5) with modifications subsequently termed (5a) and (5b). The crux of these modifications is as follows: instead of calculating randomised **G** matrices using randomised breeding values as though they were observable, directly-measured quantities like phenotypes, (5a) phenotypes are recomposed using the posterior samples of all (genetic and non-genetic) contributions to phenotype, as empirically estimated in each population.

For example, suppose that a univariate sire model is fitted in a Bayesian framework to data from each population, in order to compare simple estimates of V_a among populations (as in steps 1 and 2). In the randomisation procedure applied to each posterior sample (steps 3 and 4), individual i in population j is assigned an additive genetic value a^*_{ij} (and corresponding vectors of additive genetic values are generated for each sire's offspring according to standard rules of polygenic inheritance; Bulmer 1980, Lynch and Walsh 1998). At this point, a simulated phenotype for this individual can be recomposed according to

$$z^*_{i,j} \sim N(\mu_j + a^*_{ij}, V_{p,j} - V_{a,j})$$

where μ_j is the estimated mean phenotype, and $V_{p,j}$ and $V_{a,j}$ are the phenotypic and additive genetic variances, respectively, for population j. Having simulated these phenotypes (z^* values) for all individuals in

all populations, (5b) randomised **G** matrices are then inferred by re-fitting the same models in step (1) to the new set of recomposed phenotypes. Finally, (6) the statistics in step (2) are computed for each posterior sample of these randomised **G** matrices, generating null distributions to which the actual values from step (2) can be compared.

Application of the original and modified methods to a simple simulated example

Consider that two populations might both have identical genetic variances for some trait of interest of $V_a = 0.5$, and identical phenotypic variances of $V_p = 1.0$. Imagine that we are interested in the difference in V_a between these two populations (i.e., that in an empirical setting, we did not know that they were identical). Imagine that we estimated V_a in both populations using a standard breeding design with dams nested within sires (Lynch & Walsh 1998). This might involve taking 100 sires per population, mating each sire to five unrelated dams, and then phenotyping one offspring per dam. Data for this breeding design may be simulated according to

$$a_{i,j} \sim (0, V_a) \ ,$$

where $a_{i,j}$ is the additive genetic value of sire i in population j, and

$$z_{i,j,k} \sim (\mu_j + a_{i,j}/2, V_p - \frac{V_a}{4}) \ ,$$

where k indexes each sire's offspring, such that $z_{i,j,k}$ is the phenotype of offspring k of sire i from population j. For each population, V_a may be estimated (pretending we did not know that V_a was common between the two populations) by fitting the following linear model to simulated data

$$z_{i,j,k} = \mu_j + s_{i,j} + e_{i,j,k} \ ,$$

where the sire effects $s_{i,j}$ are treated as random, such that a variance associated with sires, $V_{sire,j}$, is estimated for each population. This sire variance would yield population-specific estimates of V_a according to $\widehat{V_{a,j}} = 4 \cdot V_{sire,j}$ (Lynch & Walsh 1998).

A valid null distribution should represent the range of values that would be obtained for a focal parameter by random chance alone, if we could replicate a given study many times in many populations with all other aspects of experimental design (e.g., sample size) held constant. We constructed a null distribution to account for all sources of statistical noise in estimating the difference in V_a, and to which we can compare ther null distributions generated both by the original Aguirre et al. (2014) method, and our modification. A valid null

distribution should represent the range of values that would be obtained for a focal parameter by random chance alone, if we could replicate a given study many times in many populations with all other aspects of experimental design (e.g., sample size) held constant. To generate such a distribution, we conducted the above simulation and analyses 1000 times, calculating the difference in V_a in each of these independent replicates of our hypothetical study. We conducted the necessary 1000 pairs of linear model analyses using restricted maximum likelihood, for computational efficiency. A method for generating a null distribution of any statistic describing differences in **G**, including one as simple as the difference between two estimates of V_a, should generater a distribution similar in breadth to the distribution generated by replicate simulations (depicted for our 1000 replicates by the dashed grey line in Fig. 1), if that method is valid.

To investigate the widths of the null distributions generated by the original Aguirre *et al.* (2014) method and our proposed modification, we took one simulation (treated as a hypothetical empirical study) and generated posterior distributions of V_a by fitting linear models using Gibbs sampling in `MCMCglmm` (Hadfield 2010). Based on this single simulation, we generated a dataset that happens to have a modest difference in estimated V_a between populations (i.e., $V_{a,2} - V_{a,1} = -0.056$). To posterior samples from this simulation, we applied both the original method (steps 1-6, as described above) and our proposed modification (steps 1-4, followed by 5a, 5b, & 6), to generate null distributions for the difference in V_a between populations.

The original method generates a null distribution of differences in V_a that is much narrower than the natural statistical variability of the system (Fig. 1a). Under the original method, there would appear to be approximately a 2.5% chance of observing a difference as large, or larger, than the test statistic of $|V_{a,2} - V_{a,1}| = 0.056$. In contrast, our proposed modification of the method generates a null distribution of differences in V_a that is similar in width to that generated under the conditions of our simulations (Fig. 1b). Under this modification, greater differences in V_a are expected, with values as large or larger than 0.056 expected 82% of the time.

Our use of REML, rather than MCMC, to fit models to the null datasets in our instructive simulations is motivated entirely by computational efficiency. We hope that this simulation (all code available in the supplement) will be useful for anyone who may wish to, for example, verify the general behaviour of either the original method, the proposed modification, or some further modifications, without having to use hundreds or thousands of CPU-hours to re-fit MCMC models. For analysis of real data, or more detailed statistical development of this type of method, one would almost certainly want to fit all models used in generating the null distribution using exactly the same methods as are applied to the real data. We compared the inference of 1000 pairs of REML mixed models and 1000 paris of MCMC mixed models to estimate differences in V_a (details and results in the supplement), and the resuting distributions were broadly similar (such that our

simple example should be entirely instructive), but not identical. As such, the approach of using REML in the generation of simulated datasets should probably not be put into empirical use, but rather, its use should be restricted to instructive numerical demonstrations. It seems possible that further development and refinement of the general approach studied here for generating null distributions of **G** matrix differences could lead to further improvements in the validity of p values, and potentially also improvements in computational efficiency. We consider further detailed study of the behaviour of the modified method, or development of more efficient versions, beyond the scope of this note.

Application to real data

We applied the original Aguirre *et al.* (2014) method, and our proposed modification, to an empirical dataset on the evolution of **G** among three Australian *Drosophila melanogaster* populations. The experimental design and laboratory procedure of the data set used here is described in detail in Hanagartner *et al.* (*submitted to Evolution simultaneously with the present manuscript*). In short, a paternal half-sibling breeding design was used to estimate additive genetic variances and covariances among sexes and four traits (desiccation resistance, cold recovery time, heat knockdown time and wing size) in each population . Heat knockdown time and desiccation resistance were measured for different individuals. Between 103 and 106 sires, and 515 and 530 full sib families, were generated for each source population, from which a total of 25068 offspring were phenotyped for one or more traits. Cold recovery time and wing size were measured for the same individuals. Separately for each population, we modelled the eight sex-specific trait measurements according to the mixed model,

$$\mathbf{z}_{ijkl} = \mathbf{X}_{ijm}\boldsymbol{\beta}_{jlm} + \mathbf{s}_{jk} + \mathbf{d}_{jl} + \mathbf{e}_{ij} \ , \tag{1}$$

where $\mathbf{z}_{ijkl}$ is the phenotype of individual i, with sex j, that is an offspring of sire k and dam l, where individual i was reared in block m. $\mathbf{X}_{ijm}$ and $\boldsymbol{\beta}_{jm}$ are the fixed effect design matrix and vector of fixed effects vector describing block effects on sex-specific trait measures. $\mathbf{s}_{jk}$ and $\mathbf{d}_{jk}$ are the random effect of sire k and dam l on traits in sex j, and $\mathbf{e}_{ij}$ are individual-level residuals for sex-specific trait measurements. Because offspring of both sexes are phenotyped for all traits, the covariances across sires are estimable for all sex-specific trait measures, and were estimated as an 8 by 8 unstructured covariance matrix, i.e.,

$$\begin{bmatrix} \mathbf{s}_{j=1,k} \\ \mathbf{s}_{j=2,k} \end{bmatrix} \sim N\left(0, \boldsymbol{\Sigma}_{\mathbf{s}}\right) \ ,$$

where $j = 1$ and $j = 2$ encode the two offspring sexes, and $\mathbf{\Sigma_s}$ is the 8 by 8 covariance matrix associated with sire identity; variances and covariances of dam effects were specified equivalently. The **G** matrix was estimated as four times the sire covariance matrix, i.e., $4\mathbf{\Sigma_s}$ (Lynch and Walsh 1998). Since (*i*) traits were treated in a sex-specific manner, and (*ii*) traits were measured in different individuals (except for cold recovery and wing size, which were measured in the same individuals in each sex), the residual covariance structure was more complex, and was specified as

$$\mathbf{e}_{ij} \sim N\left(0, \mathbf{\Sigma}_{e,j}\right) \ ,$$

where the sex-specific residual covariance matrices, $\mathbf{\Sigma}_{e,j}$, include estimated terms according to

$$\mathbf{\Sigma}_{e,j} = \begin{bmatrix} \sigma^2_{e,j,hk} & 0 & 0 & 0 \\ 0 & \sigma^2_{e,j,dr} & 0 & 0 \\ 0 & 0 & \sigma^2_{e,j,cr} & \sigma_{e,j,ws} \\ 0 & 0 & \sigma_{e,j,ws} & \sigma^2_{e,j,ws} \end{bmatrix} ,$$

where sex-specific variances and covariances are subscripted according to *hk* for heat knockdown time, *dr* for desiccation resistance, *cr* for cold recovery time, and *ws* for wing size. All traits were standardised to a variance of one and a mean of zero prior to analysis. The mixed models were implemented in `MCMCglmm` (Hadfield 2010) to obtain 1000 MCMC samples of the posterior distributions of each population-specific **G** matrix.

Here, we apply these 1000 MCMC samples to a comparison of **G** matrices among populations using a fourth-order genetic covariance tensor, as described in Aguirre *et al.* (2014). Tensors can be decomposed into a set of eigentensors ($\mathbf{E}_i$), each representing an independent component of variation among the original **G** matrices (in our case, the maximum number of non-zero components is two, one less than the number of matrices compared). We compare the observed divergence of matrices summarised by the tensor to the null distributions generated by the original Aguirre *et al.* (2014) method (steps 1-6 above), and by our proposed modification. For our modification, we applied steps (1) through (4), then (5a) and (5b), to the 1000 samples, generating 1000 randomised **G** matrices per population. Each model re-fitted in (5b) used the same burn-in and thinning interval as the original model, and generated a new set of MCMC samples that we limited to 200 to reduce computational burden (see Walter et al. 2018 for a similar approach). We constructed our null distribution of the tensor comparison from the 1000 posterior means of these latter samples.

Tensor comparisons across the populations revealed apparently significant divergence among **G** matrices when using the null distribution generated by the original method, based on non-overlap between the observed and

randomised distributions of population variation in **G** captured by the two eigentensors, $\mathbf{E}_1$ and $\mathbf{E}_2$ (Fig. 2a). In contrast, the same comparison using the modified method reveals that the divergence captured by the eigentensors falls very much within the range that would be observed if there were no differences in the **G** matrices among the populations (i.e., the estimated divergence falls well within that of the randomised distributions; Fig. 2b).

Discussion

The basic thinking underlying the Aguirre *et al.* (2014) method is likely to prove very useful in making comparisons among estimated **G** matrices. With critical modifications, we believe its application will greatly aid in making robust inferences about how patterns of genetic variation differ among populations. The modifications we suggest render the method more computationally intensive, and also may require careful consideration of nuanced issues in some key circumstances. We attempt here to begin discussion of some of these potential additional considerations.

As we demonstrate here, procedures to generate appropriate null distributions of differences among **G** matrices are likely to be more computationally intensive than the original Aguirre *et al.* (2014) method. Specifically, step (5b) of our modified method (re-fitting models to randomised data for each posterior sample of the original models) is more computationally intensive than the corresponding step (5) in the original method (calculating covariance matrices of simulated then randomised breeding values). This modification should not be seen as a detraction. Rather, it is a necessity in order to account for all sources of statistical uncertainty in estimating **G**, within the overall framework of the Aguirre *et al.* (2014) method. With the modern computational facilities of many research institutions, the need to refit models to randomised datasets based on the posterior distributions of empirical estimates of **G** should not greatly hinder the applicability of the modified method.

The modifications necessary to the Aguirre *et al.* (2014) method also reveal further subtleties associated with the details of simulating null distributions of differences among **G** matrices. In particular, step (5a) of our modified method (simulating phenotypes for each posterior sample of the original models) requires more attention than step (5) of the original method. Simple randomisation of breeding values across populations, as in the original method, will not serve the purpose for the modified steps 5a/b, because the breeding values within populations would then be independent of the population pedigrees. Rather, it seems that simulated phenotypes will generally need to be re-composed and re-analysed in each population, according to the model used to estimate **G** matrices from the real data (e.g., an analysis such as that specified by equation 1).

A different modification of the Aguirre *et al.* (2014) method has recently been suggested (Walter *et al.* 2018). By randomising phenotypes across individuals, the Walter *et al.* (2018) procedure generates a situation where phenotypes are disassociated from the pedigree and generates null distributions under the condition where **G** matrices are equal among populations because they are all null. This approach is therefore useful to generate null distributions (assuming all genetic variances and covariances were zero) for several summaries of the geometry of **G** matrices, but would not be appropriate as a null distribution to test for differences in **G** matrices among populations. A further aspect of the Walter *et al.* (2018) method that will require particular consideration in its application is that by randomising phenotypes across individuals, phenotypes will also be randomised across values of other fixed and random effects that a modern mixed model-based quantitative genetic analysis will typically consider simultaneously with inference of quantitative genetic parameters. If those other effects explain variation in the real data, and render inference of quantitative genetic parameters more precise than they would be otherwise, null distributions generated in this way will be overly pessimistic (i.e., larger than need be).

In conclusion, model-based recomposition of phenotypes under procedures closely related to the original Aguirre *et al.* (2014) method may prove useful in generating null distributions of statistics describing variation in **G** matrices that are reasonably indicative of the statistical uncertainty inherent in a given analysis. These distributions are likely to be the best representation of null hypotheses wherein genetic variation exists in multiple populations, but where there are no differences among populations. This method will compliment other recently developed approaches, which generate confidence intervals (or Bayesian credible intervals) for estimated differences (as opposed to null distributions), including Monte Carlo simulation from multivariate normal approximations to the sampling distribution of **G** matrices (Morrissey *et al.* 2012, Meyer & Houle 2013, Houle & Meyer 2015), and integration over Bayesian posterior distributions of **G** to generate credible intervals of derived quantities.

Acknowledgements

We thank C. Sgrò, T. Connallon, E. Hine, K. McGuigan, and S. Chenoweth for comments on the ideas discussed here. MBM is supported by a University Research Fellowship from the Royal Society (London). KM is supported by a Future Fellowship from the Australian Research Council.

References

Aguirre, J.D., E. Hine, K. McGuigan & M.W.Blows. 2014. Comparing **G**: multivariate analysis of genetic variation in multiple populations. Heredity 112: 21-29.

Bulmer, M.G. 1980. The Mathematical Theory of Quantititative Genetics. Clarendon Press, Oxford.

Hadfield, J. 2010. MCMC methods for multi-response generalised linear mixed models: the `MCMCglmm` `R` package. Journal of Statistical Software 33: 1-36.

Hangartner, S., C. Lasne, C.M. Sgrò, T. Connallon & K. Monro. *submitted to Evolution.* Genetic covariances promote climatic adaptation in Australian *Drosophila.*

Houle, D. and K. Meyer. 2015. Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood. Journal of Evolutionary Biology 28: 1542-1549.

Lynch, M. & B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Press, New Jersy.

Meyer, K., & D. Houle. 2013. Sampling-based approximation of confidence intervals for functions of genetic covariance matrices. Proceedings of the Association for the Advancement of Animal Breeding and Genetics 20: 523-526.

Morrissey, M.B., D.J. Parker, P. Korsten, L.E.B. Kruuk, J.M. Pemberton, T.H. Clutton-Brock, & A.J. Wilson. 2012. The prediction of adaptive evolution: empirical application of the secondary theorem of selection and comparison to the breeder's equation. Evolution 66: 2399-2410.

Walsh, B. & M. Lynch. 2018. Evolution and selection of quantitative traits. Oxford University Press, Oxford UK.

Walter, G.M., J.D. Aguirre, M.W. Blows & D. Ortiz-Barrientos. 2018. Evolution of genetic variance during adaptive radiation. The American Naturalist 191: E108-E128.

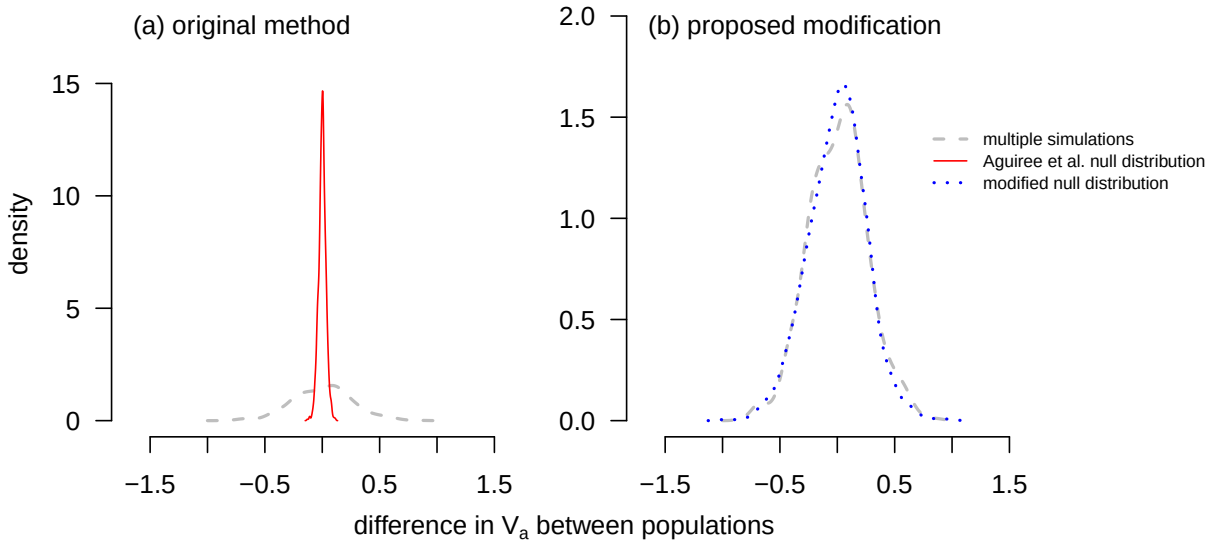

Figure 1: Comparison of (a) the original Aguirre *et al.* (2014) method, and (b) the proposed modification for generating null distributions representing statistical variability in the estimation of the difference in the additive genetic variance, V_a, between two hypothetical populations. The grey lines in both plots show the distribution of the difference in V_a frommultiple, independent, simulations in a scenario where $V_a = 0.5$ in both populations. This distribution represents the statistical uncertainty inherent in the estimation of the difference in V_a; a valid method for generating a null distribution should match this distribution. Note that the x-axis scales differ between parts (a) and (b); the distirbution of results from muitiple replicate simulations is identical on both plots.

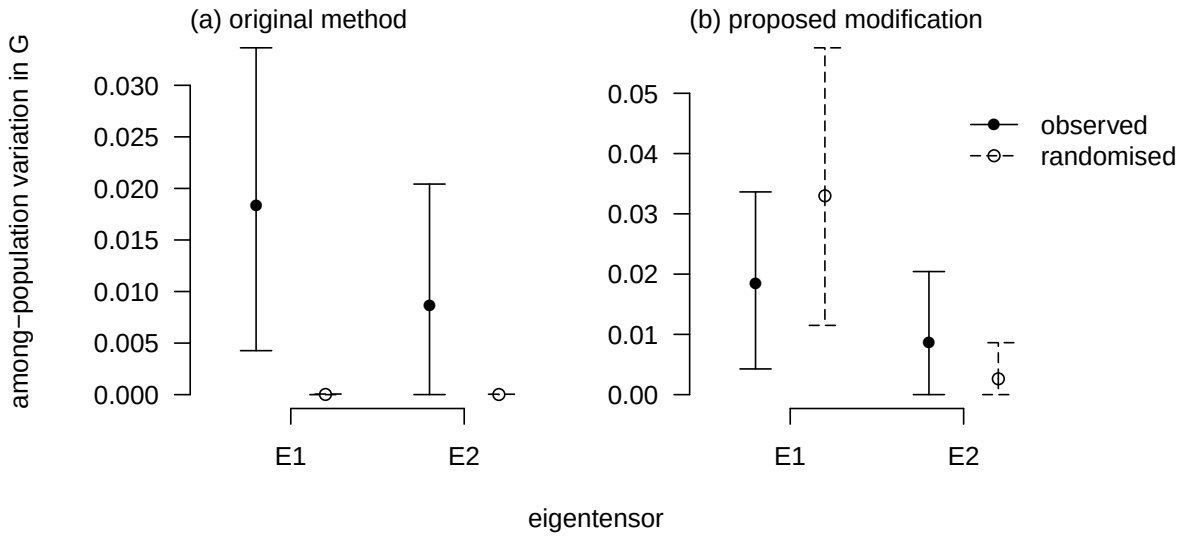

Figure 2: Results of the genetic covariance tensor analysis comparing **G** among populations using (a) the null distribution proposed by Aguirre *et al.* (2014), and (b) the proposed modification of that method. Population variation captured by the eigentensors $\mathbf{E}_1$ and $\mathbf{E}_2$ for the observed and randomized **G** are shown. Non-overlap of the intervals representing the among-population variation in **G** of the observed and randomized **G** matrices would indicate that the estimates of population divergence in **G** are larger than expected by statistical uncertainty alone.