Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 102-105

RECOVERING OBSCURED GENETIC ARCHITECTURES THROUGH DECONVOLUTION OF GWAS STATISTICS

Z. Loh

School of Environmental & Rural Science, University of New England, Armidale, NSW, 2351 Australia

SUMMARY

The aim for this study was to propose a novel deconvolution technique to improve the recovery of genetic architecture information for a trait that was obscured by strong marker linkage disequilibrium. With a simulated datasets with sample size of 3,500, this deconvolution technique has reduced the departure in distribution between true and estimated marker test statistics by up to 5.82-folds on a trait with heritability of 0.3 and 2,000 quantitative trait loci. Under the same scenario, deconvolution also improved the correlation between true and haplotype significance of association by 24.9%. Therefore, deconvolution could serve as a promising approach in recovery of genetic architecture information that could help in its elucidation in livestock traits.

INTRODUCTION

Strong linkage disequilibrium (LD) between markers has complicated the estimation of genetic architecture of various livestock traits (Lloyd-Jones *et al.* 2019). While marker pruning can be used to obtain markers that were roughly in linkage equilibrium (Chang *et al.* 2015), this method depends on arbitrarily defined parameters such as window sizes and LD thresholds, which could affect the representativeness of retained markers and thus potentially genetic architecture estimation. An alternative approach is the deconvolution of marker effect sizes using a LD matrix. This technique has been used in signal processing in other fields such as radio astronomy (Clark 1980), but studies on its use in genetic architecture parameter estimation remained sparse.

The aim for this study was to propose a novel technique for deconvolution of marker effect size and test statistics from a Genome-Wide Association Study (GWAS). This technique was tested on simulated datasets, with the aim of recovering the underlying effect size distribution and improving the correlation between true and estimated significance of association of haplotype regions.

THE DECONVOLUTION METHOD

The marker effect sizes from a GWAS $\widehat{\boldsymbol{\beta}}$ can be modelled as the result of convolution between the true effect size $\boldsymbol{\beta}$ and LD matrix $\boldsymbol{R}$ with noise $\boldsymbol{e}$ (Cheng *et al.* 2020):

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{R} * \boldsymbol{\beta} + \boldsymbol{e} \quad [1]$$

This $\boldsymbol{\beta}$ can be estimated (let this estimate be $\widetilde{\boldsymbol{\beta}}$) based on Hogböm's CLEAN algorithm. Originally introduced in radio interferometry, this algorithm iteratively removes the effects of point spread function (PSF) from the raw signals while transferring the maximums onto clean signals (Clark 1980). In this study, $\boldsymbol{R}$, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ represented the PSF, raw and clean signals respectively.

To improve the computational efficiency in cases when number of markers M exceeds sample sizes N, matrix $\boldsymbol{Z}$ (defined such that $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{R}$) was used in place of $\boldsymbol{R}$. Note that $\boldsymbol{Z}$ is not the Cholesky's factor of $\boldsymbol{R}$; $\boldsymbol{Z}$ is a rectangular matrix its j-th columns defined as follows:

$$\boldsymbol{z_j} = \frac{\boldsymbol{x_j} - 2p_j}{\sqrt{\sum(\boldsymbol{x_j} - 2p_j)^2}} \quad [2]$$

where $\boldsymbol{X}$ is the raw genotype array, and p_j is the allele frequency for marker j.

Let $\hat{\boldsymbol{t}}$ be the raw test statistics from GWAS. While other definitions such as Aguilar *et al.* (2019)

can be used with this method, for this study the $\hat{\boldsymbol{t}}$ was defined as $\widehat{\boldsymbol{\beta}}$ scaled by its standard deviation s (Gondro 2015). The deconvolution starts by identifying the genome-wide top marker in term of $|\hat{\boldsymbol{t}}|$ (let m be this marker), from which the corresponding marker effect size $\hat{\beta}_m$ and m-th column of $\boldsymbol{Z}$, $\boldsymbol{z_m}$, were obtained. This marker effect size was recorded as the m-th elements of $\widetilde{\boldsymbol{\beta}}$. To prevent overcorrection of marker effect sizes which could destabilize the algorithm, the m-th element of $\widehat{\boldsymbol{\beta}}$ was "muted" by setting it as "NAN" and no longer used in any subsequent iterations. The effect sizes and test statistics unexplained by $\hat{\beta}_m$, $\widehat{\boldsymbol{\beta}}_{(1)}$ and $\hat{\boldsymbol{t}}_{(1)}$, were estimated as follows (Gondro 2015):

$$\widehat{\boldsymbol{\beta}}_{(1)} = \widehat{\boldsymbol{\beta}} - \hat{\beta}_m \boldsymbol{Z}^T \boldsymbol{z}_m \quad [3]$$

$$\hat{\boldsymbol{t}}_{(1)} = \frac{\widehat{\boldsymbol{\beta}}_{(1)}}{s} \quad [4]$$

with the subscript (k) denotes the k-th iteration of this algorithm. The next iteration starts by identifying the top marker in term of $|\hat{\boldsymbol{t}}_{(1)}|$ and its corresponding $\hat{\beta}_m$ and $\boldsymbol{z}_m$. This process iterates until all markers were deconvolved. The deconvolved test statistics $\tilde{\boldsymbol{t}}$ could be obtained by scaling $\widetilde{\boldsymbol{\beta}}$ with s as in [4], as empirical simulations suggested that $\widetilde{\boldsymbol{\beta}}$ scaled with s better reflects the true QTL effect sizes as it considers the inflation of estimated marker effect sizes in the $\widehat{\boldsymbol{\beta}}$.

TESTING THE DECONVOLUTION METHOD

The deconvolution method was tested using Python (version 3.11.5, released 11 September 2023) with simulated genotypes and phenotypes under changing parameter values, with the parameter tested and their associated values provided in Table 1.

Table 1. Parameters and their associated values tested in this experiment

Parameters	Default Value	Alternative Value
Sample size	3500	1500
Number of QTL	2000	500
Genotype Array	$\boldsymbol{X_{2K}}$	$\boldsymbol{X_{10K}}$

Two 60k genotype arrays (denoted as $\boldsymbol{X_{2K}}$ and $\boldsymbol{X_{10K}}$ respectively) were simulated using QMSim (Sargolzaei and Schenkel 2009) with historical population size set at 2,000 and 10,000 respectively, and were gene-dropped for 12,000 and 60,000 generations respectively. 20 chromosomes of 100 cM and mutation rate at 2.5×10^{-5} were used for both arrays, and sample sizes as provided in Table 1.

Within a genotype array $\boldsymbol{X}$, 2000 or 500 markers were designated as QTL and their effect size $\boldsymbol{\beta}$ simulated using gamma distribution $gamma(0.4\,, 1)$. Phenotype $\boldsymbol{y}$ was then calculated as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

where $\boldsymbol{e}$ is the residual component simulated using normal distribution $N\left(0\,, \frac{(1-h^2)var(\boldsymbol{X\beta})}{h^2}\right)$ with the heritability h^2 set at 0.3. The GBLUP-backsolved marker effect sizes $\widehat{\boldsymbol{\beta}}$ and test statistics $\hat{\boldsymbol{t}}$ were obtained as described by VanRaden (2008) and Gondro (2015). The expected test statistics with zero LD between QTL and estimation error, $\boldsymbol{t}$, were also calculated from $\boldsymbol{\beta}$ as $t_i = \beta_i \sqrt{(N-2)var(\boldsymbol{x_i})} / \sqrt{var(\boldsymbol{y}) - \beta_i^2 var(\boldsymbol{x_i})}$ (Wang and Xu, 2019). Deconvolution was then applied onto $\widehat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{t}}$ to yield the corresponding $\widetilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{t}}$. Minor allele frequency filtering of 0.05 was applied on GWAS and the deconvolution algorithm. The deconvolved test statistics $\tilde{\boldsymbol{t}}$ (denoted as DCV) were tested against raw test statistics $\hat{\boldsymbol{t}}$ (denoted as RAW), as well as LD pruning and equal distance pruning methods (denoted as LDT and EQL respectively). For LDT, the markers were pruned as in "indep-pairwise" from PLINK 2.0 (Chang *et al.* 2015) with window size, step size and LD threshold at 50, 10 and 0.4 respectively. For EQL, the fifth of every non-overlapping 10 SNPs windows were chosen.

Two metrics were used to measure the methods' performance: departure in distribution (DiD) and correlation between true and marker haplotype scores (r_h^2). The DiD was measured with Wasserstein's distance between the distribution of $|\boldsymbol{t}|$ and $|\tilde{\boldsymbol{t}}|$ as defined by Vasershtein (1969), with a lower statistic indicates a reduced departure and improved performance. To ensure the retained markers accurately represent the haplotypes' significance of association, the DiD was calculated using the top marker within a marker window, with the marker being used as the window's midpoint. To maintain the genotyping density, markers that were removed during the pruning process were assigned as zero in $\tilde{\boldsymbol{t}}$. For r_h^2, the correlation between the sums of $|\boldsymbol{t}|$ within each haplotype blocks with that from raw marker $|\hat{\boldsymbol{t}}|$ or deconvolved $|\tilde{\boldsymbol{t}}|$ were obtained (Villiers *et al.* 2024). The blocks were defined as the window of markers in the DiD, with both EQL and LDT methods being tested.

This experiment was repeated 40 times for each set of parameter values. Welch's t-test was used to test the significance of differences in performance between methods and parameters, with two tests considered significantly different if the logarithmically transformed p-values $logpval > 2$.

RESULTS AND DISCUSSION

Compared to other methods tested, deconvolution significantly reduces departure in distributions DiD between true and marker test statistics (Table 2); for example, under default parameter values, deconvolution reduces the DiD by 3.22-folds from 0.954 to 0.296, which was also 2.86-folds and 1.15-folds decrease compared to LDT and EQL respectively ($logpval$ up to 114.8 between DCV and RAW) (Figure 1). This differences in performance were even more significant in $\boldsymbol{X_{10K}}$, achieving a 5.82-folds reduction from 0.972 in RAW to 0.167 in DCV ($logpval = 127.7$). Reducing the sample size from 3500 to 1500 also reduces the DiD for DCV from 0.297 to 0.221 ($logpval = 73.6$). This however contrasted with the significantly increased DiD for all other methods ($logpval$ up to 51.5 for RAW and 39.1 for LDT). These results highlighted the negative impacts of LD between markers in obscuring a trait's QTL effect size distribution especially with small sample sizes, and that the effects of LD can be removed using deconvolution.

Table 2. Departure in distribution (DiD) between true and raw (RAW) test statistics, test statistics pruned with LD threshold (LDT) and by 10 markers window (EQL) and from deconvolution (DCV). Superscripts with different letters denote significant differences between methods

Parameters	Values	Method			
		RAW	LDT	EQL	DCV
Sample size (default)	3500	0.954^{a}	0.849^{b}	0.341^{c}	0.296^{d}
(alternative)	1500	0.975^{a}	0.868^{b}	0.356^{c}	0.221^{d}
Number of QTL	500	0.953^{a}	0.849^{b}	0.347^{c}	0.295^{d}
Genotype array	$\boldsymbol{X_{10K}}$	0.972^{a}	0.957^{b}	0.369^{c}	0.167^{d}

Deconvolution also significantly increases the correlations of haplotype scores between true and estimated haplotype scores r_h^2 for both LD threshold pruning (LDT) and 10 marker windows (EQL) haplotype block definitions (Table 3); with the exception of default parameter values with EQL where the improvement of DCV from RAW was not deemed statistically significant ($logpval = 1.27$); significant improvements were observed for all other parameter values with both EQL and LDT methods. These improvements were even more significant for $\boldsymbol{X_{10K}}$ and in the oligogenic case of 500 QTL; for example, deconvolution improves the correlation from 0.650 to 0.825 in LDT ($logpval = 16.1$) and from 0.748 to 0.843 in EQL ($logpval = 8.04$) in the case with 500 QTL, a more significant improvement than with 2,000 QTL, suggesting deconvolution improves how well the marker haplotype score in reflecting the true contribution toward the phenotype.

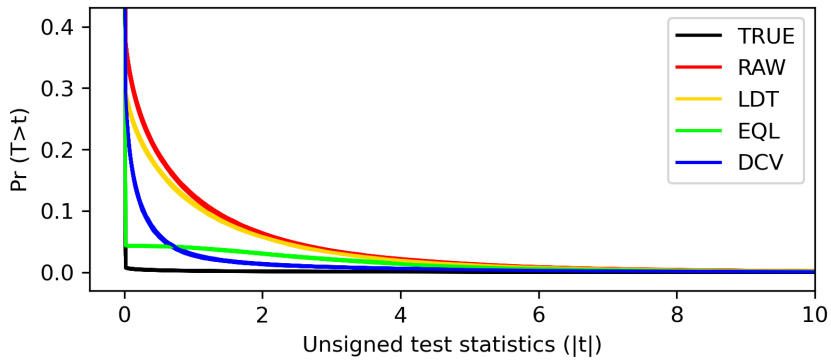

Figure 1. Top 40% tail distribution of true (TRUE), raw (RAW), LD threshold-pruned (LDT), 10 marker window (EQL) and deconvolved (DCV) test statistics under default scenario

Table 3. Correlation in haplotype score for raw (RAW) and deconvolved (DCV) test statistics, r_h^2, for haplotype block by LD threshold or by 10 markers window under changing parameter values. Superscripts with different letters denote significant differences between methods

Parameters	Values	LD threshold pruned (LDT)		10 marker windows (EQL)	
		RAW	DCV	RAW	DCV
Sample size (default)	3500	0.457[a]	0.578[b]	0.571[b]	0.610[b]
(alternative)	1500	0.253[a]	0.351[c]	0.328[b]	0.391[c]
Number of QTL	500	0.650[a]	0.825[c]	0.748[b]	0.843[c]
Genotype array	$\boldsymbol{X_{10K}}$	0.288[a]	0.431[c]	0.332[b]	0.444[c]

CONCLUSION

A novel technique for deconvolution of GWAS statistics that improves the recovery of genetic architecture information was proposed. This technique has successfully reduced the departure in distribution between true and inferred effect size distributions and improved how well the marker haplotype score reflects the true contribution under varying scenarios. It is anticipated this method could be used on a real datasets to elucidate the genetic architectures for various livestock traits.

REFERENCES

Aguilar I., Legarra A., Cardoso F., Masuda Y., Lourenco D. and Misztal I. (2019) *Genet. Sel. Evol.* **51**: 28.

Chang C.C., Chow C.C., Tellier L.C.A.M., Vattikuti S., Purcell S.M. and Lee J.J. (2015) *GigaScience* **4**: 7.

Cheng W., Ramachandran S. and Crawford L. (2020) *PLOS Genet.* **16**: e1008855.

Clark B.G. (1980) *Astron. Astrophys.* **39**: 377.

Gondro C. (2015) 'Primer to analysis of genomic data using R'. Springer, Cham.

Lloyd-Jones L.R., Zeng J., Sidorenko J., Yengo L., Moser G., Kemper K.E., Wang H., Zheng Z., Magi R., Esko T., Metspalu A., Wray N.R., Goddard M.E., Yang J. and Visscher P.M. (2019) *Nat. Commun.* **10**: 5086.

Sargolzaei M. and Schenkel F.S. (2009) *Bioinformatics* **25**: 680.

VanRaden P.M. (2008) *J. Dairy Sci.* **75**: 3136.

Vasershtein L.N. (1969) *Probl. Inf. Transm.* **5**: 64.

Villiers K., Voss-Fels K.P., Dinglasan E., Jacobs B., Hickey L. and Hayes B.J. (2024) *Plant Genome-US* **17**: e20467.

Wang M. and Xu S. (2019) *Heredity* **123**: 287.