Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models

Running Head: Structural equation modeling for association studies

Mehdi Momen[1], Ahmad Ayatollahi Mehrgardi[1*], Mahmoud Amiri Roudbar [1], Andreas Kranis[2], Renan Mercuri Pinto[3,4], Bruno D. Valente[4], Gota Morota[5,] Guilherme J. M. Rosa [4,6], Daniel Gianola[4,6,7]

[1] Department of Animal Science, Faculty of Agriculture, Shahid Bahonar University of Kerman (SBUK), Kerman, Iran
[2] Roslin Institute, University of Edinburgh, Midlothian, UK, EH25 9PS
[3] Department of Exact Sciences, University of São Paulo - ESALQ, Piracicaba-SP, Brazil
[4] Department of Animal Sciences, University of Wisconsin, Madison, WI, USA
[5] Department of Animal Science, University of Nebraska-Lincoln, Lincoln, Nebraska, USA
[6] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA
[7] Department of Dairy Science, University of Wisconsin, Madison, WI, USA

[*]Corresponding author: mehrgardi@uk.ac.ir

Email addresses:

MM: momenmehdi@yahoo.com

AAM: mehrgardi@uk.ac.ir

MA: mahmoud.amiri225@gmail.com

GM: morota@unl.edu

AK: andreas.kranis@roslin.ed.ac.uk

RMP: rpinto@wisc.edu

BDV: bvalente@wisc.edu

GJMR: grosa@wisc.edu

DG: gianola@ansci.wisc.edu

Abstract

Background

Phenotypic networks describing putative causal relationship among multiple phenotypes can be used to infer single-nucleotide polymorphism (SNP) effects in genome-wide association studies (GWAS). In GWAS with multiple phenotypes, reconstructing underlying causal structures among traits and SNPs using a single statistical framework is essential for understanding the entirety of genotype-phenotype maps. A structural equation model (SEM) can be used for such purpose.

Methods

We applied SEM to GWAS (SEM-GWAS) in chickens, taking into account putative causal relationships among body weight (BW), breast meat (BW), hen-house production (HHP), and SNPs. We assessed the performance of SEM-GWAS by comparing the model results with those obtained from traditional multi-trait association analyses (MTM-GWAS).

Results

Three different putative causal path diagrams were inferred from highest posterior density (HPD) intervals of 0.75, 0.85, and 0.95 using the IC algorithm. A positive path coefficient was estimated for BM→BW, and negative values were obtained for BM→HHP and BW→HHP in all implemented scenarios. Further, the application of SEM-GWAS enabled decomposition of SNP effects into direct, indirect, and total effects, identifying whether a SNP effect is acting directly or indirectly on given trait. In contrast, MTM-GWAS only captured overall genetic effects on traits, which is equivalent to combining the direct and indirect SNP effects from SEM-GWAS.

Conclusions

Our results suggested that SEM-GWAS provides insights into mechanisms by which SNPs affect traits through partitioning effects into direct, indirect, and total components. Thus, we provide evidence that SEM-GWAS captures complex relationships and delivers a more comprehensive understanding of SNP effects compared to MTM-GWAS.

Key words: Causal structure, GWAS, multiple traits, path analysis, SEM, SNP effect

Background

Genome-wide association studies (GWAS) have become a standard approach for investigating relationships between common genetic variants in the genome (e.g., single-nucleotide polymorphisms, SNPs) and phenotypes of interest in human, plant, and animal genetics [1-3]. A typical GWAS is based on univariate linear or logistic regression of phenotypes on genotypes for each SNP individually while often adjusting for the presence of nuisance covariates [4]. A statistically significant association indicates that SNPs may be in strong linkage disequilibrium (LD) with quantitative trait loci (QTLs) that contribute to the trait etiology. Alternatively, multi-trait model GWAS (MTM-GWAS) can be used to test for genetic associations among a set of traits [5-7]. It has been established that MTM-GWAS reduces false positives and increases the statistical power of association tests, explaining the recent popularity of this method. MTM-GWAS can be used to study genetic associations of multiple traits; however, it does not identify factors that mediate relationships between the detected effects and dependencies involving complex traits.

Complex traits are the product of various cryptic biological signals that may affect a trait of interest either directly or indirectly through other intermediate traits [8]. A standard regression cannot describe such complex relationships between traits and QTLs properly. For instance, some traits may simultaneously act as both dependent and independent variables. Structural equation modeling (SEM) is an extended version of Wright's path analysis [9, 10], that offers a powerful technique for modeling causal networks. In a complex genotype-phenotype setting involving many traits, a given trait can be influenced not only by genetic and systematic factors but also by other traits (as covariates) as well. Here, QTLs may not affect the target trait directly; instead, the effects may be mediated by upstream traits in a causal network. Indirect effects may therefore constitute a proportion of perceived pleiotropy, and these concepts apply to sets of heritable traits, organized as networks, are common in biological systems. An example from dairy cattle production systems, described by Gianola and Sorensen [10] and Rosa, et al. [11], is that higher milk yield increases the risk of a particular disease, such as mastitis, while the prevalence of the disease may negatively affect milk yield. In humans, obesity is a key factor influencing insulin resistance, which subsequently causes type 2 diabetes. A list of causal networks across human diseases and candidate genes is described in Kumar and Agrawal [12] and Schadt [13].

Although MTM-GWAS is a valuable approach, it only captures correlations or associations among traits and does not provide information about causal relationships. Knowledge of the causal structures underlying complex traits is essential, as correlation does not imply causation. For example, a correlation between two traits, T1 and T2, could be attributed to a direct effect of T1 on T2, T2 on T1, or to additional variables that jointly influence both traits [11]. Likewise, if we know a "causal" SNP is linked to a QTL, we can imagine three possible scenarios: 1) causal

95 ($SNP \rightarrow T1 \rightarrow T2$), 2) reactive ($SNP \rightarrow T2 \rightarrow T1$), or 3) independent ($T1 \leftarrow SNP \rightarrow T2$).
96 Scenarios (1) and (2), do not causes pleiotropy but produce association.
97 A SEM methodology has the ability to handle complex genotype-phenotype maps in GWAS
98 placing an emphasis on causal networks [14]. Therefore, SEM-based GWAS (SEM-GWAS) may
99 provide a better understanding of biological mechanisms and of relationships among a set of
100 traits than MTM-GWAS. SEM can potentially decompose the total SNP effect on a trait into
101 direct and indirect (i.e., mediated) contributions. However, SEM-derived GWAS has not been
102 discussed or applied fully in quantitative genetic studies yet. Our objective was to illustrate the
103 potential utility of SEM-GWAS by using three production traits in broiler chickens genotyped for
104 a battery of SNP as a case example.

105 # Methods

106 ## Data set

107 The analysis included records for 1,351 broiler chickens provided by Aviagen Ltd. (Newbridge,
108 Scotland) for three phenotypic traits: body weight (BW), ultrasound of breast muscle (BM) at 35
109 days of age, and hen-house egg production (HHP), defined as the total number of eggs laid
110 between weeks 28 and 54 per bird. The sample consisted of 274 full-sib families, 326 sires, and
111 592 dams. More details regarding population and family structure were provided by Momen et al.
112 [15]. A pre-correction procedure was performed on the phenotypes to account for systematic
113 effects such as sex, hatch week, pen, and contemporary group for BW, BM, and HHP.

114 Each bird was genotyped for 580,954 SNP markers with a 600k Affymetrix SNP [16] chip
115 (Affymetrix, Inc., Santa Clara, CA, USA). The Beagle software [17] was used to impute missing
116 SNP genotypes, and quality control was performed using PLINK version 1.9 [18]. After

117 removing markers that did not fulfill the criteria of minor allele frequencies $< 1\%$, call
118 rate $> 95\%$, and Hardy–Weinberg equilibrium (Chi-square test p-value threshold was 10^{-6}),
119 354,364 autosomal SNP markers were included in the analysis.

120 ## Multiple-trait model for GWAS

121 MTM-GWAS is a single-trait GWAS model extended to multi-dimensional responses. When
122 only considering additive effects of SNPs, the phenotype of a quantitative trait using the single-
123 trait model can be described as:

124 $$y_i = \sum_{q=1}^{k} x_{iq}\beta_q + w_{ij}s_j + e_i \qquad (1)$$

125 where y_i is the phenotypic trait of individual *i*, x_{iq} is the incidence value for the *i*th phenotype in
126 the *q*th level of systematic environmental effect, β_q is fixed effect of the *q*th systemic
127 environmental effect on the trait, $w_j = (w_1, \dots, w_p)$ is the number of A alleles (i.e., $w_j \in \{0,1,2\}$)
128 in the genotype of SNP marker *j*, and s_j is the allele substitution effect for SNP marker *j*. Strong
129 LD between markers and QTLs coupled with an adequate marker density increases the chance of
130 detecting marker and phenotype associations. Hypothesis testing is typically used to evaluate the
131 strength of the evidence of a putative association. Typically, a *t*-test is applied to obtain p-values,
132 and the statistic is $T_{ij} = \hat{s}_j / se(\hat{s}_j)$, where $\hat{s}$ is the point estimate of the *j*th SNP effect and $se(\hat{s}_j)$
133 is its standard error.

134 The single locus model described above is naïve for a complex trait because the data typically
135 contain hidden population structure and individuals have varying degrees of genetic similarity
136 [19, 20]. Therefore, accounting for covariance structure induced by genetic similarity is expected

to produce better inferences [21]. Ignoring effects that reveal genetic relatedness inflates the residual terms, compromises the ability to detect association. A random effect g_i including a covariance matrix reflecting pairwise similarities between additive genetic effects of individuals can be included to control population stratification. The similarity metrics can be derived from pedigree information or from whole-genome marker genotypes. This model extended for analysis of t traits is given by:

$$y_{il} = \sum_{q=1}^{k} x_{iq}\beta_{ql} + w_{ij}s_{jl} + g_{il} + e_{il} \qquad (2)$$

for $i = 1,2,\cdots,n$, $l = 1,2,\cdots,t$. In this extension, y_{il} is the phenotypic value of the lth trait for the ith subject, β_{qj} is the systematic effect of the qth environmental factor x_{iq} on the lth trait, s_{jl} is the additive effect of the jth marker on the lth trait, w_{ij} is as previously defined, and g_{il} and e_{il} are the random polygenic effect and model residual assigned to individual i for trait l, respectively. Random effects within a trait follow the multivariate normal distribution, $\begin{bmatrix}\boldsymbol{g}_l \\ \boldsymbol{e}_l\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}\sigma^2_{g_l} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}\sigma^2_{e_l}\end{bmatrix}\right)$, where $\boldsymbol{K}$ is genetic relationship matrix, $\sigma^2_{g_l}$ is the additive genetic variance of trait l, $\boldsymbol{I}$ is an identity matrix, and $\sigma^2_{e_l}$ is the residual variance for trait l. The multiple-trait model accounts for the additive genetic ($\rho_{ll'}$) and residual correlation ($\lambda_{ll'}$) between a pair of traits l and l'.

The positive definite matrix **K** may be a genomic relationship matrix (**G**) computed from marker data, or a pedigree-based matrix (**A**) computed from genealogical information. The **A** matrix describes the expected additive similarity among individuals, while the **G** measures the realized

fraction of alleles shared. Genomic relationship matrices can be derived in several ways [22-24]. Here, we used the form proposed by VANRADEN 2008:

$$\mathbf{G} = \frac{\mathbf{MM}'}{2\sum p_j q_j} \qquad (3)$$

where $\mathbf{M}$ is an $n \times p$ matrix of centered SNP genotypes and p_j and $q_j = 1 - p_j$ are the allele frequencies at marker locus j. We evaluated both **A** and **G** in the present study.

Structural equation model association analysis

A SEM consists of two essential parts: a measurement model and a structural model. The measurement model depicts the connections between observable variables and their corresponding latent variables. The measurement model is also known as confirmatory factor analysis. The critical part of a SEM is the structural model, which can have three forms. The first consists of observable exogenous and endogenous variables. This model is a restricted version of a SEM known as path analysis [9]. The second form explains the relationship between exogenous and endogenous variables that are only latent. The third type is a model consisting of both manifest and latent variables.

SEM can be applied to GWAS as an alternative to MTM-GWAS to study how different causal paths mediate SNP effects on each trait. The following SEM model was considered:

$$y_{il} = \mu_l + \sum_{m \in C} y_m \boldsymbol{\lambda}_{lm} + w_{j(l)} s_{j(l)} + g_{il} + \varepsilon_{il} \qquad (4)$$

where C is the set of phenotypic traits that directly affect the trait l, $\boldsymbol{\lambda}_{lm}$is a structural coefficient representing the effect of trait m on trait l, and $g_l \sim N(0, \boldsymbol{K}\sigma_l^2)$ is the polygenic effect of the lth trait. The remaining terms are as presented earlier with one important difference: the SNP effects

are not interpreted as overall effects on trait l but instead represent direct effects on trait l. Additional indirect effects from the same SNP may be mediated by phenotypic traits in *C*. Each marker is entered into equation (4) one at a time, and its significance is tested. For a discussion of how SEM represents genetic signals on each trait through multiple causal paths, see Valente, et al. [25]. Despite the difference in interpretation, the distribution of the vector of polygenic effects is assumed to be the same as in the MTM-GWAS model. The same applies to residual terms within a trait. We also consider trait-specific residuals to be independent within an individual. This restriction is required to render structural coefficients likelihood-identifiable. In addition, the interpretation of inferences as having a causal meaning requires imposing the restriction that the residuals' joint distribution be interpreted as the causal sufficiency assumption [26]. In the present study, all exogenous and endogenous variables were observable, and there was no latent variable. In hence, causal structure was assumed between the endogenous variables BM, BW, and HHP.

We considered the following GWAS models, which their causal structures were recovered by the inductive causation (IC) algorithm [26]: (1) MTM-GWAS with pedigree-based kinship **A** (MTM-A) or marker-based kinship **G** (MTM-G), and (2) SEM-GWAS with **A** (SEM-A) or **G** (SEM-G). Although nuisance covariates such as environmental factors can be omitted in the graph, they may be incorporated into the models as exogenous variables. The SEM representation allowed us to decompose SNP effects into direct, indirect, and total effects.

A direct SNP effect is the path coefficient between a SNP as an exogenous variable and a dependent variable without any causal mediation by any other variable. The indirect effects of a SNP are those mediated by at least one other intervening endogenous variable. Indirect effects are calculated by multiplying path coefficients for each path linking the SNP to associated variable,

and then summing over all such paths [9]. The overall effect is the sum of all direct and indirect effects. By explicitly accounting for complex relationship structure among traits in such a way, SEM provides a better understanding of a genome-wide SNP analysis by allowing us to decompose effects into direct, indirect, and overall effects within a predefined casual framework.

MTM-GWAS and SEM-GWAS were compared with the logarithm of the likelihood function (log L), Akaike's Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The model providing the lowest values for these information criteria is considered to fit the data better [27]. MTM-GWAS and SEM-GWAS were fitted using the SNP Snappy strategy, which is implemented in the Wombat software program [28].

Searching for a phenotypic causal network in a mixed model

In the SEM-GWAS formulation described earlier, the structure of the underlying causal phenotypic network needs to be known. Because this is not so in practice, we used a causal inference algorithm to infer the structure. Residuals are assumed to be independent in all SEM analyses, so associations between observed traits are viewed as due to causal links between traits and by correlations among genetic values (i.e., g_1, g_2, and g_3). Thus, to eliminate confounding problem when inferring the underlying network among traits, we used the approach of Valente, et al. [29] to search for acyclic causal structures through conditional independencies on the distribution of the phenotypes, given the genetic effects. A causal phenotypic network was inferred in two stages: 1) a MTM model [30] was employed to estimate covariance matrices of additive genetic effects and of residuals, and 2) the causal structure among phenotypes from the covariance matrix between traits, conditionally on additive genetic effects inferred by the IC algorithm. The residual (co)variance matrix was inferred using Bayesian MCMC [29, 31], with samples drawn from the posterior distribution. For each query testing statistical independence

between traits $y_l$and $y_{l'}$, the posterior distribution of the residual partial correlation $\rho_{y_l,y_{l'}}|S$ was obtained, where S is a set of variable (traits) that are independent. Three highest posterior density (HPD) intervals of 0.75, 0.85, and 0.95 were used to make statistical decisions for SEM-GWAS. We thus considered SEM-A75 (HPD > 0.75), SEM-A85 (HPD > 0.85), SEM-A95 (HPD > 0.95), and SEM-G75 (HPD > 0.75). An HPD interval that does not contain zero declares y_l and $y_{l'}$ to be conditionally dependent.

Results

Figure 1 shows phenotypic relationship structures recovered by the IC algorithm for the three different HPD intervals. Edges connecting two traits represent non-null partial correlations as indicated by HPD intervals. We compared the two MTM-GWAS and four SEM-GWAS by using the three chicken traits (BW, BM, and HHP). Only causal structures among the three traits are shown in Figure 1, because other parts were the same across the different SEM models. Fully recursive SEM-A75 and SEM-G75 revealed direct effects of BM on BW and HHP, and those of BW on HHP, as well as an indirect effect of BM on HHP. In addition, SEM-A85 detected a direct effect of BM on BW, the direct effect of BW on HHP, and the indirect effect of BM on HHP mediated by BW. Finally, SEM-A95 only identified a direct effect of BM on BW because of a statistically stringent HPD cutoff imposed.

Given the causal structures inferred from the IC algorithm, the following SEM was fitted:

$$\begin{cases} \boldsymbol{y}_1 = \mu + \boldsymbol{Z}_i\boldsymbol{g}_1 + W_{ij}S_j + \boldsymbol{\varepsilon}_i \\ \boldsymbol{y}_2 = \mu + \lambda_{21}\boldsymbol{y}_1 + \boldsymbol{Z}_i\boldsymbol{g}_2 + W_{ij}S_j + \boldsymbol{\varepsilon}_i \\ \boldsymbol{y}_3 = \mu + \lambda_{31}\boldsymbol{y_1} + \lambda_{32}\boldsymbol{y}_2 + \boldsymbol{Z}_i\boldsymbol{g}_3 + W_{ij}S_j + \boldsymbol{\varepsilon}_i \end{cases} \quad (5)$$

Note that only a small number of the entries in the structural coefficient matrix (λ in equation 5) are nonzero due to sparsity. These nonzero entries specify the effect of one phenotype on other

phenotypes. The corresponding directed acyclic graph is shown in Figure 2 assuming the causal relationships among the three traits, where y_1, y_2, and y_3 represent BM, BW, and HHP, respectively; SNP_j is the genotype of the *j*th SNP; S_{jl} is the direct SNP effect on trait l; and the remaining variables are as presented earlier. This diagram depicts a fully recursive structure in which all recursive relationships among the three phenotypic traits are shown. Arrows represent causal connections, whereas double-headed arrows between polygenic effects are correlations.

<< Figure 1 about here>>

<< Figure 2 about here>>

We examined the fit of each model implemented to assess how well it describes the data (Table 1). Valente, et al. [25] showed that re-parametrization and reduction of a SEM mixed model yield the same joint probability distribution of observation as in MTM suggesting that expected likelihood of SEM and MTM should be the same. As expected, SEM-GWAS and MTM-GWAS showed very similar results (e.g., SEM-A75 vs. MTM-A and SEM-G75 vs. MTM-G). Among the models considered, the ones involving **G** exhibited a better fit. SEM-A85 and SEM-A95, sharing a subset of the SEM-A75 structure, presented almost identical AIC and BIC values.

<<Table 1 about here>>

Structural coefficients

Table 2 presents the causal structural path coefficients for endogenous variables (BM, BW, and HHP). All models have positive effects for BM→BW, whereas the BM→HHP and BW→HHP relationships have negative path coefficients. The latter confirmed the fact that chicken breeding is divided into broiler and layer sections due to the negative genetic correlation between BW and HHP.

<<**Table 2 about here**>>

Also shown in Table 2 are the magnitudes of the SEM structural coefficient reflecting the intensity of the causality. The positive coefficient λ_{21} quantifies the (direct) causal effect of BM on BW. This suggests that a 1-unit increase in BM results in a λ_{21} -unit increases in BW. Likewise, the negative causal effects λ_{31} and λ_{32} offer the same interpretation.

Decomposition of SNP effect paths using a fully recursive model

We can decompose the SNP effects into direct and indirect effects using Figure 2. The direct effect of the SNP j on y_3 (HHP) is given by $d_{SNP_j \to y_3}$: $\hat{S}_{j(y3)}$, where d denotes the direct effect. Note there are only one direct and many indirect paths. We find three indirect paths from SNP_j to y_3 mediated by y_1 and y_2 (i.e., the nodes formed by other traits). The first indirect effect is $ind_{(1)SNP_j \to y_3}$: $\lambda_{32}(\lambda_{21}\hat{S}_{j(y1)})$ in the path mediated by y_1 and y_2, where ind denotes the indirect effect. The second indirect effect $ind_{(2)SNP_j \to y_3}$: $\lambda_{32}\hat{S}_{j(y2)}$, is mediated by y_2. The last indirect effect, is $ind_{(3)SNP_j \to y_3}$: $\lambda_{31}\hat{S}_{j(y1)}$, mediated by y_1. Therefore, the overall effect is given by summing all four paths, $T_{SNP_j \to y_3}$: $\lambda_{32}(\lambda_{21}\hat{S}_{j(y1)}) + \lambda_{32}\hat{S}_{j(y2)} + \lambda_{31}\hat{S}_{j(y1)} + \hat{S}_{j(y3)}$. The fully recursive model of the overall SNP effect is then:

$$\begin{cases} T_{\hat{S}_{j\to y_1}: \hat{S}_{j(y1)}} \\ T_{\hat{S}_{j\to y_2}: \lambda_{21}(\hat{S}_{j(y1)})+\hat{S}_{j(y2)}} \\ T_{\hat{S}_{j\to y_3}: \lambda_{32}[\lambda_{21}(\hat{S}_{j(y1)})+\hat{S}_{j(y2)}]+\lambda_{31}(\hat{S}_{j(y1)})+\hat{S}_{j(y3)}} \end{cases} \quad (6)$$

For y_1 (BM), there is only one effect, so the overall effect is equal to the direct effect. For y_2 (BW) and y_3 (HHP), direct and indirect SNP effects are involved. There are two paths for y_2: one indirect, $ind_{S_j \to y_2}$: $\hat{S}_{j(y1)} \to y_1 \to y_2$, and one direct, $d_{S_j \to y_2}$: $\hat{S}_{j(y2)} \to y_2$. Here, SNP effect is

direct and mediated thorough other phenotypes according to causal networks in SEM-GWAS (Figures 1 and 2). For instance, the overall SNP effect for y_3 into four direct and indirect paths is $T_{\hat{S}_{j\to y_3}}: \lambda_{32}\lambda_{21}\hat{S}_{j(y1)} + \lambda_{32}\hat{S}_{j(y1)} + \lambda_{31}\hat{S}_{j(y1)} + \hat{S}_{j(y3)}$.

The scatter plots in Figure 3 compare the estimated total effects for HHP ($T_{\hat{S}_{j\to y_3}}$) obtaind from SEM-GWAS and those from MTM-GWAS. We observed good agreement between SEM-GWAS and MTM-GWAS. The total SNP signals derived from SEM and MTM are the same but SEM provides biologically relevant additional information.

<<Figure 3 about here>>

Supplementary Figures S1-S4 present scatter plots of MTM-GWAS and SEM-GWAS signals (SEM-A75, SEM-G75, SEM-A85, and SEM-A95) for the $BM \to BW$ path, which was a common path across all SEM-GWAS considered. These two traits have a genetic correlation of 0.5 (results not shown). We break the SEM causal link into direct, indirect, and overall effects based on the IC algorithm with HPD > 0.85, whereas MTM-GWAS capture an overall SNP effect on BW. Scatter plots of the overall effects from SEM-GWAS and the total effects from MTM-GWAS indicated almost perfect agreement (top left plots, Supplementary Figures S1–S4). We observed concomitance between estimated overall and direct effects (top right plots, Supplementary Figures S1–S4). In contrast, there was less agreement in the magnitude of the SNP effects when comparing overall vs. indirect effects (bottom left plots, Supplementary Figures S1–S4). There was no linear relationship between the indirect and direct SNP effects (bottom right plots, Supplementary Figures S1–S4). In short, genetic signals detected in SEM-GWAS were close to those of MTM-GWAS for overall effects because both models are based on a multivariate

307 approach with the same covariance matrix. In all SEM-GWAS, results showed that direct effect
308 contributed to overall effects than the indirect effects.

309 ## Manhattan plot of direct, indirect, and overall SNP effects

310 Figure 4 depicts a Manhattan plot summarizing the magnitude of direct, indirect, and overall SNP
311 effects. We plotted the decomposed SNP effects on BW along chromosomes to visualize
312 estimated marker effects from SEM-GWAS. The indirect and direct effects provide a view of
313 SNP effects from a perspective that is not available for the total effect of MTM-GWAS. For
314 instance, many pleiotropic QTLs have positive direct effects on BW but negative effects on BM.
315 There were two estimated SNP effects on chromosomes 1 and 2 that deserve particular attention.
316 These two SNPs are highlighted with black circles and red ovals. The overall effect of the first
317 SNP consisted of large indirect and small direct effects on BM, whereas the opposite pattern was
318 observed for the second SNP, which showed large direct and small indirect effects. Although the
319 overall effects of these SNPs were similar (top Manhattan plot, Figure 4), use of decomposition
320 allowed us to find out that the trait of interest is affected in different manners: the second SNP
321 effect acted directly on BW without any mediation by BM, whereas the first SNP reflected a
322 large effect mediated by BM on BW. Collectively, new insight regarding the direction of SNP
323 effects can be obtained using the SEM-GWAS methodology.

324 It should also be noted that the estimated additive SNP effects obtained from the four SEM-
325 GWAS can be used for inferring pleiotropy. For instance, a pleiotropic QTL may have a large
326 positive direct effect on BW but may exhibit a negative indirect effect coming from BM, which
327 in turn reduces the total QTL effect on BW. Arguably, the methodology employed here would be
328 most effective when the direct and indirect effects of a QTL are in opposite directions. If the

329 direct and indirect QTL effects are in the same direction, the power of SEM-GWAS may be the
330 same as the overall power of MTM-GWAS.

331 **<<Figure 4 about here>>**

332 Discussion

333 It is becoming increasingly common to analyze a set of traits simultaneously in GWAS by
334 leveraging genetic correlations between traits [32, 33]. In the present study, we illustrated the
335 potential utility of a SEM-based GWAS approach, which has the potential advantage of
336 embedding a pre-inferred causal structure across phenotypic traits [29]. SEM-GWAS accounts
337 for the relationship of mediating variables that could be either dependent or independent with
338 restriction on a residual covariance. This is a useful approach when multiple mediators interplay
339 influencing the final outcomes [34, 35]. SEM-GWAS is achieved by first inferring the structure
340 of network between phenotypic traits. For this purpose, we used a modified version of the IC
341 algorithm described by Valente, et al. [29]. The IC algorithm was used to explore putative causal
342 links among phenotypes obtained from a residual covariance matrix, in a model that accounted
343 for systematic and genetic confounding factors such as polygenic additive effects. It then
344 produced a posterior distribution of partial residual correlations between any possible pairs of
345 variables. Three different causal path diagrams were inferred from HPD intervals of 0.75, 0.85,
346 and 0.95. We observed that the number of identified paths decreased with an increase in the HPD
347 interval value. Only a path connecting BM and BW was present in all HPD intervals considered.
348 Moreover, we found that the partial residual correlation between BM and HHP was weaker than
349 that between BM and BW. This may explain why the path between BM and HHP was not
350 detected with HPD intervals larger than 0.75.

Estimated path coefficients reflect the strength of each causal link. For instance, a positive path coefficient from BM to BW suggests that a unit increase in BM directly results in an increase in BW. Our results showed that MTM-GWAS and SEM-GWAS were similar in terms of the goodness of fit as per the AIC and BIC criteria. This finding is in agreement with theoretical work of Valente, et al. [25] showing the equivalence between models. Thus, MTM-GWAS and SEM-GWAS produced the same marginal phenotypic distributions and goodness of fit values. A similar approach has been proposed by Li, et al. [14], Mi, et al. [36], and Wang and van Eeuwijk [37]. The main difference between our approach and theirs is that they used SEM in the context of standard QTL mapping, whereas our SEM-GWAS is developed for GWAS based on a linear mixed model.

The advantage of SEM-GWAS over MTM-GWAS is that the former decomposes SNP effects by tracing inferred causal networks. Our results showed that by partitioning SNP effect into direct, indirect, and total components, an alternative perspective of SNP effects can be obtained. As shown in Figure 4, direct and indirect effects may differ in magnitude and sign, acting in the same direction or even antagonistic manners. Note that the total SNP effects inferred from SEM-GWAS were the same as the estimated SNP effects from MT-GWAS (Figure 3). However, knowledge derived from the decomposition of SNP effects may be critical for animal and plant breeders to breaking unfavorable indirect QTL effects, or to obtain better SNP effects estimates than those from MTM-GWAS [e.g., 36].

Conclusion

SEM offers insights into how phenotypic traits relate to each other. We illustrated potential advantages of SEM-GWAS relative to the commonly used standard MTM-GWAS by using three

chicken traits as an example. SNP effects pertaining to SEM-GWAS have a different meaning than those in MTM-GWAS. Our results showed that SEM-GWAS enabled the identification of whether a SNP effect is acting directly or indirectly, i.e. mediated, on given trait. In contrast, MTM-GWAS only captures overall genetic effects on traits, which is equivalent to combining direct and indirect SNP effects from SEM-GWAS together. Thus, SEM-GWAS offers more information and provides an alternative view of putative causal network, enabling a better understanding the genetic quiddity of traits at the genomic level.

Conflict of Interest

The authors do not have any conflict of interest.

Author's contributions

MM carried out the study and wrote the first draft of the manuscript. GJMR and DG designed the experiment, supervised the study and critically contributed to the final version of manuscript. GM contributed to the interpretation of results, provided critical insights, and revised the manuscript. BDV and AAM participated in discussion and reviewed the manuscript. MA, AK and RMP contributed materials and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgment

The first author wishes to acknowledge the Ministry of Science, Research and Technology of Iran for financially supporting his visit to the University of Wisconsin-Madison. Work was partially

393 supported by the Wisconsin Agriculture Experiment Station under hatch grant 142-PRJ63CV to
394 DG.

395 References

396 1. Visscher Peter m, Brown Matthew a, Mccarthy Mark i, Yang J, Five Years of GWAS
397 Discovery, The American Journal of Human Genetics. 2012; 90: 7-24.
398 2. Hayes B, Goddard M, Genome-wide association and genomic selection in animal breeding
399 This article is one of a selection of papers from the conference "Exploiting Genome-wide
400 Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for
401 sustainable farming", Genome. 2010; 53: 876-883.
402 3. Brachi B, Morris G P, Borevitz J O, Genome-wide association studies in plants: the missing
403 heritability is in the field, Genome Biology. 2011; 12: 232.
404 4. Sikorska K, Lesaffre E, Groenen P F, Eilers P H, GWAS on your notebook: fast semi-parallel
405 linear and logistic regression for genome-wide association studies, BMC Bioinformatics.
406 2013; 14: 1-11.
407 5. Zhou X, Stephens M, Genome-wide efficient mixed-model analysis for association studies,
408 Nat Genet. 2012; 44: 821-824.
409 6. Korte A, Vilhjálmsson B J, Segura V, Platt A, Long Q, Nordborg M, A mixed-model approach
410 for genome-wide association studies of correlated traits in structured populations, Nature
411 genetics. 2012; 44: 1066-1071.
412 7. O'reilly P F, Hoggart C J, Pomyen Y, Calboli F C, Elliott P, Jarvelin M-R, Coin L J,
413 MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS, PloS one.
414 2012; 7: e34861.
415 8. Falconer D S, Mackay T F, Introduction to quantitative genetics (4th edn), Trends in Genetics.
416 1996; 12: 280.
417 9. Wright S, Correlation and causation, Journal of agricultural research. 1921; 20: 557-585.
418 10. Gianola D, Sorensen D, Quantitative genetic models for describing simultaneous and
419 recursive relationships between phenotypes, Genetics. 2004; 167: 1407-1424.

11. Rosa G J, Valente B D, De Los Campos G, Wu X-L, Gianola D, Silva M A, Inferring causal phenotype networks using structural equation models, Genetics Selection Evolution. 2011; 43: 6.

12. Kumar S, Agrawal S, Disease-oriented Causal Networks, Encyclopedia of Systems Biology, Springer, 2013, pp. 593-594.

13. Schadt E E, Chapter 10 - Reconstructing Causal Network Models of Human Disease A2 - Lehner, Thomas, in: Miller B L, State M W (Eds.), Genomics, Circuits, and Pathways in Clinical Neuropsychiatry, Academic Press, San Diego, 2016, pp. 141-160.

14. Li R, Tsaih S-W, Shockley K, Stylianou I M, Wergedal J, Paigen B, Churchill G A, Structural model analysis of multiple quantitative traits, PLoS genetics. 2006; 2: e114.

15. Momen M, Mehrgardi A A, Sheikhy A, Esmailizadeh A, Fozi M A, Kranis A, Valente B D, Rosa G J M, Gianola D, A predictive assessment of genetic correlations between traits in chickens using markers, Genetics Selection Evolution. 2017; 49: 16.

16. Kranis A, Gheyas A A, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P, Development of a high density 600K SNP genotyping array for chicken, BMC genomics. 2013; 14: 59.

17. Browning S R, Browning B L, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, Am J Hum Genet. 2007; 81.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A, Bender D, Maller J, Sklar P, De Bakker P I, Daly M J, PLINK: a tool set for whole-genome association and population-based linkage analyses, The American Journal of Human Genetics. 2007; 81: 559-575.

19. Listgarten J, Lippert C, Kadie C M, Davidson R I, Eskin E, Heckerman D, Improved linear mixed models for genome-wide association studies, Nat Meth. 2012; 9: 525-526.

20. Gianola D, Fariello M I, Naya H, Schön C-C, Genome-Wide Association Studies with a Genomic Relationship Matrix: A Case Study with Wheat and Arabidopsis, G3: Genes, Genomes, Genetics. 2016; 6: 3241-3256.

21. Galesloot T E, Van Steen K, Kiemeney L A, Janss L L, Vermeulen S H, A comparison of multivariate genome-wide association methods, PloS one. 2014; 9: e95923.

22. Vanraden P, Efficient methods to compute genomic predictions, J Dairy Sci. 2008; 91: 4414 - 4423.

451 23. Yang J, Benyamin B, Mcevoy B, Gordon S, Henders A, Nyholt D, Common SNPs explain a
452 large proportion of the heritability for human height, Nat Genet. 2010; 42: 565 - 569.
453 24. Forni S, Aguilar I, Misztal I, Different genomic relationship matrices for single-step analysis
454 using phenotypic, pedigree and genomic information, Genet Sel Evol. 2011; 43: 1.
455 25. Valente B D, Rosa G J, Gianola D, Wu X-L, Weigel K, Is structural equation modeling
456 advantageous for the genetic improvement of multiple traits?, Genetics. 2013; 194: 561-572.
457 26. Pearl J, Causal inference in statistics: An overview, Statistics surveys. 2009; 3: 96-146.
458 27. Akaike H, Information theory and an extension of the maximum likelihood principle,
459 Selected Papers of Hirotugu Akaike, Springer, 1998, pp. 199-213.
460 28. Meyer K, Tier B, Churchill G A, "SNP Snappy": A Strategy for Fast Genome-Wide
461 Association Studies Fitting a Full Mixed Model, Genetics. 2012; 190: 275-277.
462 29. Valente B D, Rosa G J, De Los Campos G, Gianola D, Silva M A, Searching for recursive
463 causal structures in multivariate quantitative genetics mixed models, Genetics. 2010; 185:
464 633-644.
465 30. Henderson C, Quaas R, Multiple trait evaluation using relatives' records, Journal of Animal
466 Science. 1976; 43: 1188-1197.
467 31. Wu X L, Heringstad B, Gianola D, Bayesian structural equation models for inferring
468 relationships between phenotypes: a review of methodology, identifiability, and applications,
469 Journal of Animal Breeding and Genetics. 2010; 127: 3-15.
470 32. Gao H, Zhang T, Wu Y, Wu Y, Jiang L, Zhan J, Li J, Yang R, Multiple-trait genome-wide
471 association study based on principal component analysis for residual covariance matrix,
472 Heredity. 2014; 113: 526-532.
473 33. Wu B, Pankow J S, Genome-wide association test of multiple continuous traits using imputed
474 SNPs, Statistics and its interface. 2017; 10: 379.
475 34. Bellavia A, Valeri L, Decomposition of the total effect in the presence of multiple mediators
476 and interactions, American journal of epidemiology. 2017.
477 35. Barfield R, Shen J, Just A C, Vokonas P S, Schwartz J, Baccarelli A A, Vanderweele T J, Lin
478 X, Testing for the indirect effect under the null for genome-wide mediation analyses, Genetic
479 epidemiology. 2017; 41: 824-833.

36. Mi X, Eskridge K, Wang D, Baenziger P S, Campbell B T, Gill K S, Dweikat I, Bovaird J, Regression-based multi-trait QTL mapping using a structural equation model, Statistical applications in genetics and molecular biology. 2010; 9.

37. Wang H, Van Eeuwijk F A, A New Method to Infer Causal Phenotype Networks Using QTL and Phenotypic Information, PLOS ONE. 2014; 9: e103997.

Figures

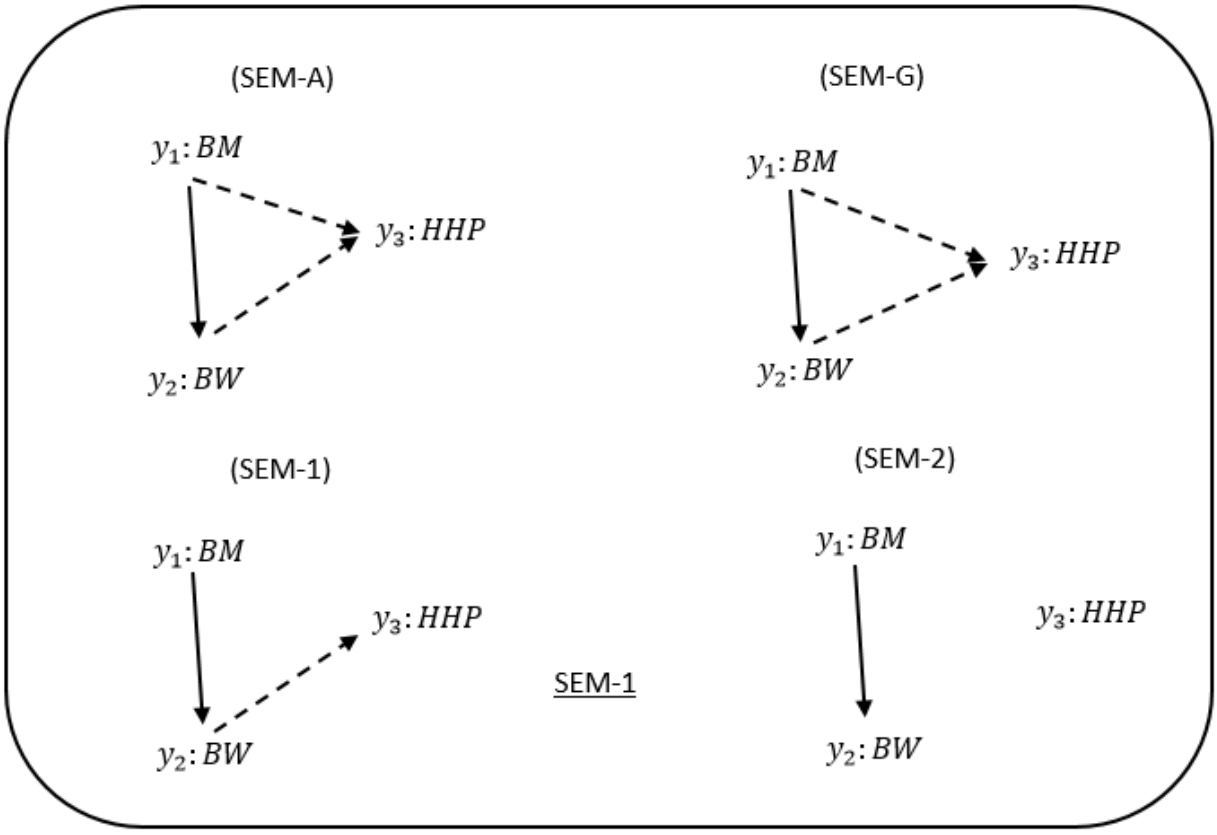


Figure 1 Causal graphs inferred using the IC algorithm among three traits: breast meat (BM), body weight (BW) and hen-house production (HHP) in the chicken data. SEM-A75 and SEM-G75 were the inferred fully recursive causal structures with HPD > 0.75 and corrected for genetic confounder using **A** (pedigree-based) and **G** (marker-based) matrices. SEM-A85 and SEM-A95 were obtained with HPD > 0.85 and HPD > 0.95, respectively, corrected with **A**. Arrows indicate direction of causal relationships. Dashed lines indicate negative coefficients, and the continuous arrows indicate positive coefficients.

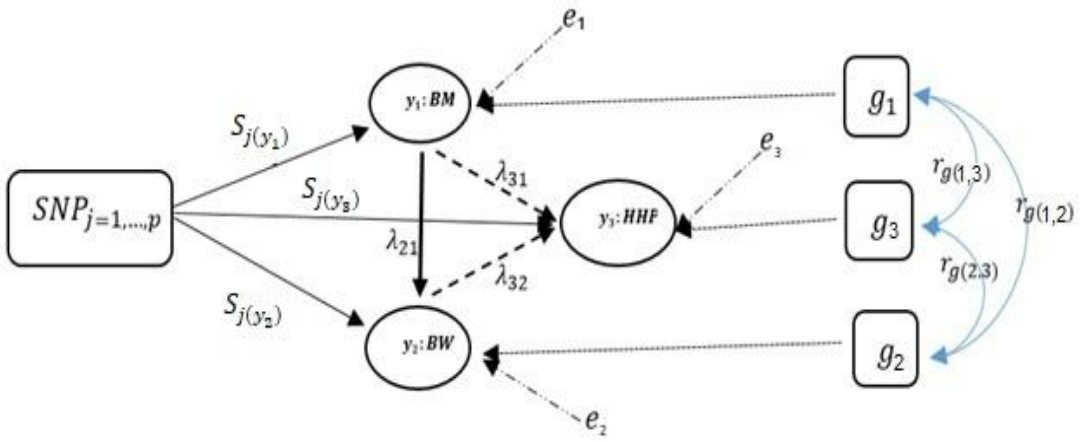


Figure 2 A diagram for causal path analysis of SNP effects in a fully recursive structural equation model for three traits, p exogenous independent SNP variables, and three correlated polygenic effects. Arrows indicate the direction of causal effects and dashed lines represent associations among the three phenotypes. Genetic correlation between traits (r_g), polygenic effects (g_l), environmental effect on trait l (e_l), effects of j th SNP on l th trait ($S_{j(y_l)}$), and recursive effect of phenotype l' on phenotype l ($\lambda_{l,l'}$).

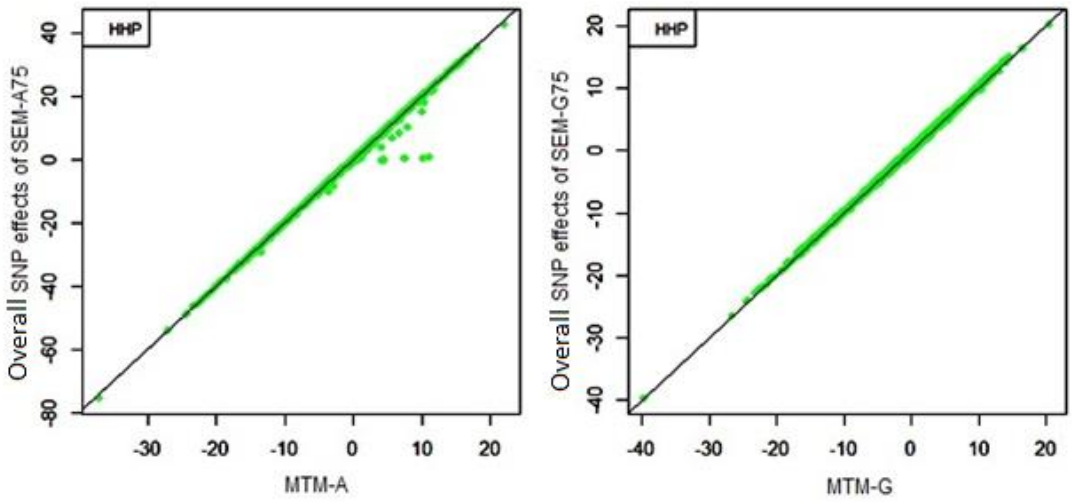
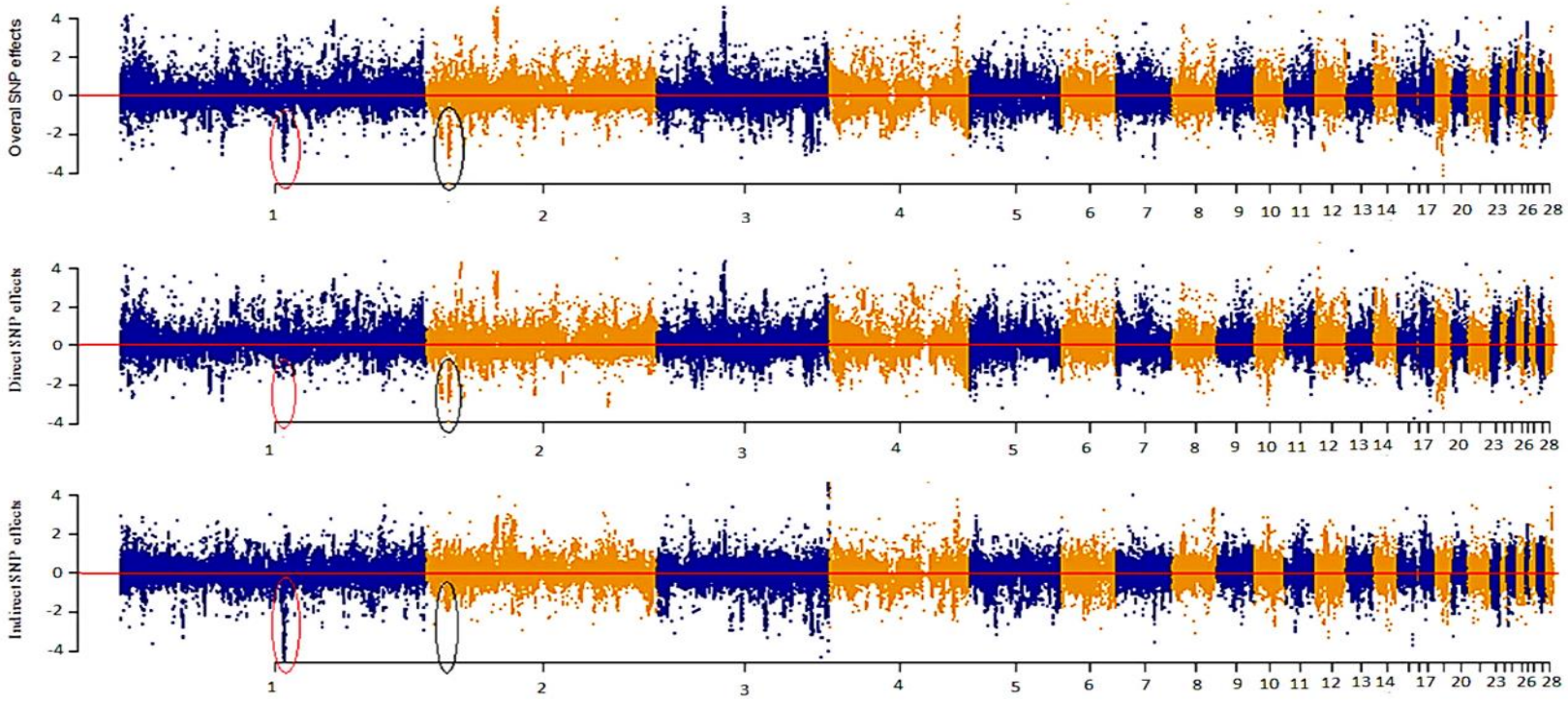


Figure 3 Comparison of multiple trait (MTM) and fully recursive overall SNP effects obtained with A (pedigree-based) and G (marker-based) from structural equation modeling (SEM)-based GWAS. Overall effects in SEM are the sum of all direct and indirect effects. HHP: hen-house egg production.



560

561 **Figure 4 Manhattan plot showing overall, direct, and indirect SNP effects using a full recursive model based on G matrix for**
562 **body weight (BW).**

563

Tables

Table 1 Model comparison criteria: logarithm of the restricted maximum likelihood function (log L), Akaike's information criteria (AIC), Schwarz Bayesian information criteria (BIC) to evaluate model fit for two MTM and four SEM models.

Model	Maximum log L	-1/2 AIC	-1/2 BIC
MTM-A	-7093.480	-7105.48	-7142.436
SEM-A75	-7098.370	-7110.415	-7147.321
SEM-A85	-7095.188	-7107.188	-7144.143
SEM-A95	-7097.517	-7109.517	-7146.470
MTM-G	-6529.270	-6541.276	-6578.232
SEM-G75	-6537.391	-6549.391	-6586.34

A: pedigree-based relationship matix, G: VanRaden's mrker-based relationship matrix

575

Table 2 Estimates of three causal structural coefficients (λ) derived from four different structural models. BM: breast meat. BW: body weight. HHP: hen-house production. SEM-75: HPD > 0.75. SEM-G75: HPD > 0.75. SEM-A85: HPD > 0.85. SEM-A95: HPD > 0.95.

Path	Structural Models			
	SEM-75	SEM-G75	SEM-A85	SEM-A95
$\lambda_{BM\rightarrow BW}(\lambda_{21})$	2.13	2.19	2.14	2.14
$\lambda_{BM\rightarrow HHP}(\lambda_{31})$	-0.17	-0.28	***	***
$\lambda_{BW\rightarrow HHP}(\lambda_{32})$	-0.27	-0.096	-0.31	***

576

577

578

579

580

581

582

583