Molecular Genetics and Genomics manuscript No.
(will be inserted by the editor)

Detecting genetic epistasis by differential departure from independence

**Ruby Sharma · Zeinab Sadeghian Tehrani · Sajal Kumar ·
Mingzhou Song**

Received: date / Accepted: date

Abstract Countering prior beliefs that epistasis is rare, genomics advancements suggest the other way. Current practice often filters out genomic loci with low variant counts before detecting epistasis. We argue that this practice is far from optimal because it can throw away strong epistatic patterns. Instead, we present the compensated Sharma-Song test to infer genetic epistasis in genome-wide association studies by differential departure from independence. The test does not require a minimum number of replicates for each variant. We also introduce algorithms to simulate epistatic patterns that differentially depart from independence. Using two simulators, the test performed comparably to the original Sharma-Song test when variant frequencies at a locus are marginally uniform; encouragingly, it has a marked advantage over alternatives when variant frequencies are marginally nonuniform. The test further revealed uniquely clean epistatic variants associated with chicken abdominal fat content that are not prioritized by other methods. Genes involved in most numbers of inferred epistasis between single nucleotide polymorphisms (SNPs) belong to pathways known for obesity regulation; many top SNPs are located on chromosome 20 and in intergenic regions. Measuring differential departure from independence, the compensated Sharma-Song test offers a practical choice for studying epistasis robust to nonuniform genetic variant frequencies.

Keywords Differential departure from independence · Epistasis · Chicken obesity · Genome-wide association study

1 Introduction

More than one genetic variant can be associated with disease such as asthma (Howard et al., 2002), diabetes (Cho et al., 2004), and schizophrenia (Andreasen et al., 2012). Epistasis refers to two or more genetic variants having a collaborative phenotypic effect different from a simple combination of their independent effects. Epistatic genotype-phenotype association is increasingly revealed in populations via high-throughput combinatorial genome screening (Domingo et al., 2019). However, statistical methods to discover epistatic interactions in genome-wide association studies (GWAS) still seem discordant on how to quantify an epistatic effect (Cordell, 2002; Domingo et al., 2019).

Many statistical measures of epistasis are available (Jing and Shen, 2015; Wan et al., 2010; Ueki and Cordell, 2012; Purcell et al., 2007; Tuo et al., 2017; Zhang and Liu, 2007; Yang et al., 2009). Parametric methods for epistasis often

Ruby Sharma
Department of Computer Science, New Mexico State University, Las Cruces, New Mexico, United States
E-mail: ruby49@nmsu.edu

Zeinab Sadeghian Tehrani
MS Program in Bioinformatics, Department of Computer Science, New Mexico State University, Las Cruces, New Mexico, United States
E-mail: zeinabs@nmsu.edu

Sajal Kumar
Department of Computer Science, New Mexico State University, Las Cruces, New Mexico, United States
E-mail: sajal49@nmsu.edu

Mingzhou Song
Department of Computer Science, Molecular Biology and Interdisciplinary Life Sciences Graduate Program, New Mexico State University, Las Cruces, New Mexico, United States
E-mail: joemsong@cs.nmsu.edu

use logistic regression, including the EPISNPmpi (Ma et al., 2008), allele-based test (Purcell et al., 2007), LASSO (Fan and Li, 2001), SCAD (Winham et al., 2011) and the joint-effects method (Ueki and Cordell, 2012). These methods can be subject to parametric model biases (Niel et al., 2015). Non-parametric methods such as BOOST (Wan et al., 2010), carrying minimal model biases, must overcome type I errors due to data unevenness, where the marginal distribution of variants at a genomic locus can be highly nonuniform. Many methods skip candidate variants without a minimum number of replicates for each variant (Niel et al., 2015), closely related to the minor allele frequency. Such unevenness violates the Cochran condition (Cochran, 1952, 1954) in the context of Pearson's chi-squared test. Although this practice can indeed remove spurious epistatic patterns, it can also miss strong epistatic patterns as illustrated in Discussion.

To address the challenges, we study epistasis by differential departure from independence, in contrast to differential joint distribution. The rationale is that the joint distribution of two mutations can differ due to only marginal differences, where a pair independently affects a trait, not a truly collaborative effect. Here, we present the compensated Sharma-Song test to detect genetic epistasis in GWAS by differential departure from independence. The test is robust to data unevenness without requiring a minimum number of replicates for each variant. It extends the original Sharma-Song test (Sharma et al., 2021) that detects second-order different patterns across contingency tables. We also introduce algorithms to simulate discrete patterns driven by differential departure from independence.

We benchmarked the compensated test against popular epistasis inference methods including the allele-based test (Purcell et al., 2007), BOOST (Wan et al., 2010) and the joint-effects test (Ueki and Cordell, 2012), all implemented in open-source software Plink (Purcell et al., 2007). The data were synthesized using both our simulator and GAMATES (Urbanowicz et al., 2012) display-no-marginal-effects (DNME) and display-marginal-effects (DME) models (Shang et al., 2011; Xie et al., 2012; Jing and Shen, 2015). The compensated test performed close to the original test when data are even; it, however, has a marked advantage over other methods when data are uneven.

Finally, we apply all methods on pairwise interactions among 48,034 single nucleotide polymorphisms (SNPs) in a genome-wide association study on 475 male chickens (Guo et al., 2011). The compensated test reports about 99 million epistatic pairs significantly associated with chicken fat content. Mutations, pathways, and chromosomes involved with detected epistatic patterns are supported by literature for their functional effects on obesity. The top epistatic patterns determined by each method are notably different, with the compensated test returning the cleanest additive by additive epistatic SNP-SNP patterns. Measuring differential departure from independence, the compensated Sharma-Song test offers a practical choice for studying genetic epistasis while being robust to nonuniform variant frequencies.

2 Methods

2.1 Epistasis by differential departure from independence

To characterize pattern differences across conditions, we define first-, second-, and full-order differential patterns that are first introduced by Sharma et al. (2021). Let X and Y be two discrete random variables of r and s levels, respectively. For condition $k \in \{1,\ldots,K\}$, let $p_k(X,Y)$ be the joint probability mass function of X and Y; let $p_k(X)$ and $p_k(Y)$ be the marginal distributions of X and Y. The K patterns are *conserved* if

$$p_1(X,Y) = \cdots = p_K(X,Y) \tag{1}$$

Otherwise, they are *differential*. The K patterns are *first-order conserved* if

$$p_1(X) = \cdots = p_K(X) \text{ and } p_1(Y) = \cdots = p_K(Y) \tag{2}$$

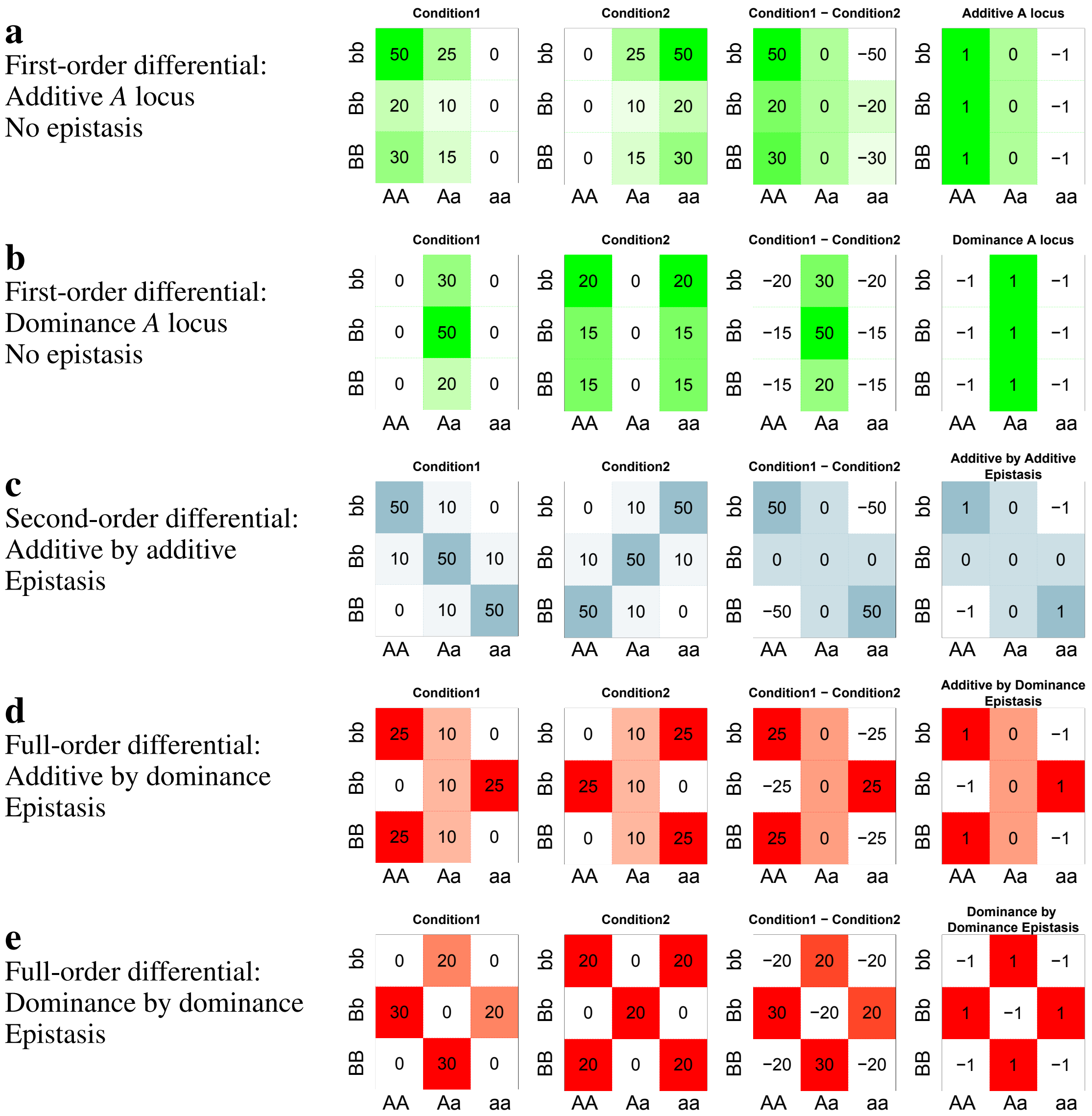
Otherwise, they are *first-order differential*. The K patterns are *second-order conserved* if

$$p_1(X,Y) - p_1(X)p_1(Y) = \cdots = p_K(X,Y) - p_K(X)p_K(Y) \tag{3}$$

Otherwise, they are *second-order differential* or have *differential departure from independence*. Patterns are *full-order differential* if they are both first- and second-order differential.

First-, second- and full-order differential patterns can be linked to five epistasis types (Kempthorne, 1957) as shown in Figure 1. Each pattern is represented by a 3×3 contingency table formed by two genomic loci A and B, each composed of diploid alleles. The columns represent genotype AA, Aa, aa; the rows represent genotype BB, Ba, bb. The count in each entry of the table is the number of subjects with corresponding locus A (column) and locus B (row) genotypes, known as genotype frequency. Figure 1a,b are first-order patterns where the two loci are independent but of different marginal frequencies across conditions. The loci are not epistatic and exhibit single-locus additive or dominance effect. In the second-order differential patterns in Fig. 1c, the two loci have differential departure from independence but same marginal frequencies between conditions, representing additive-by-additive epistatic effects. Figure 1d,e are full-order patterns having differential departure from independence with different marginal frequencies across conditions; the patterns represent the additive-by-dominance and dominance-by-dominance epistatic effects, respectively.

Second-order patterns are related to DNME epistasis models, whereas full-order patterns are related to DME epistasis models (Shang et al., 2011; Xie et al., 2012; Jing and Shen, 2015).



	Condition1			Condition2			Condition1 − Condition2			Additive A locus		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
bb	50	25	0	0	25	50	50	0	−50	1	0	−1
Bb	20	10	0	0	10	20	20	0	−20	1	0	−1
BB	30	15	0	0	15	30	30	0	−30	1	0	−1

	Condition1			Condition2			Condition1 − Condition2			Dominance A locus		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
bb	0	30	0	20	0	20	−20	30	−20	−1	1	−1
Bb	0	50	0	15	0	15	−15	50	−15	−1	1	−1
BB	0	20	0	15	0	15	−15	20	−15	−1	1	−1

	Condition1			Condition2			Condition1 − Condition2			Additive by Additive Epistasis		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
bb	50	10	0	0	10	50	50	0	−50	1	0	−1
Bb	10	50	10	10	50	10	0	0	0	0	0	0
BB	0	10	50	50	10	0	−50	0	50	−1	0	1

	Condition1			Condition2			Condition1 − Condition2			Additive by Dominance Epistasis		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
bb	25	10	0	0	10	25	25	0	−25	1	0	−1
Bb	0	10	25	25	10	0	−25	0	25	−1	0	1
BB	25	10	0	0	10	25	25	0	−25	1	0	−1

	Condition1			Condition2			Condition1 − Condition2			Dominance by Dominance Epistasis		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
bb	0	20	0	20	0	20	−20	20	−20	−1	1	−1
Bb	30	0	20	0	20	0	30	−20	20	1	−1	1
BB	0	30	0	20	0	20	−20	30	−20	−1	1	−1

Fig. 1 Linking first-, second- and full-order differential patterns to non-epistasis and epistasis types. The first two columns are genotype frequency tables of two loci *A* (column) and *B* (row) in two conditions. The third column is the subtraction of the first table by the second table in the same sub-figure. The last column is the matching epistasis pattern. The color intensity in a table entry is proportional to counts in that entry. **(a,b)** Two first-order only differential patterns, mapping to additive or dominance *A* locus, respectively, a single-variant effect without epistasis. **(c)** A second-order only differential pattern, mapping to additive by additive epistasis. **(d,e)** Two full-order differential patterns, mapping to additive by dominance and dominance by dominance epistasis, respectively.

2.2 The compensated Sharma-Song test

The compensated Sharma-Song test detects second-order differential patterns and handles nonuniform marginal distributions, a common issue in contingency-table tests like Pearson's chi-squared test. It is an extension of original Sharma-Song test (Sharma et al., 2021). The compensated test examines contingency tables $\mathbf{C}_1, \ldots, \mathbf{C}_K$ of the same dimensions $r \times s$

and detects 2nd-order differential patterns. It *compensates* contingency table $\mathbf{C}_k = \left[n_{ij}^k\right]$ by uniformly adding an equal fraction of a sample to every entry in the table:

$$n_{ij}^k \leftarrow n_{ij}^k + \frac{1}{rs}, \quad i = 1, \ldots, r; j = 1, \ldots, s \tag{4}$$

Next, it follows the same steps in the original Sharma-Song test. We first quantify normalized departure from independence $\mathbf{A}_k = \left[a_{ij}^k\right]$ by

$$a_{ij}^k = \frac{n_{ij}^k - \bar{n}_{ij}^k}{\sqrt{\bar{n}_{ij}^k}}, \quad \text{where } \bar{n}_{ij}^k = \frac{\sum_{i'=1}^r n_{i'j}^k \cdot \sum_{j'=1}^s n_{ij'}^k}{\sum_{i'=1}^r \sum_{j'=1}^s n_{i'j'}^k} \tag{5}$$

$\bar{n}_{ij}^k$ represents expected count for entry (i, j) in $\mathbf{C}_k$ if row and column variables are independent. We convert matrix $\mathbf{A}_k$ to vector $\mathbf{e}_k$ of dimension $(r-1)(s-1)$ via the Helmert transform. Under the null hypothesis, components in $\mathbf{e}_k$ are i.i.d. standard normal. Measuring deviation of each vector $\mathbf{e}_k$ to their weighted average, the vectors Δ_k form columns in matrix $\mathbf{Q}$. Rows of $\mathbf{Q}$ are projected to $\mathbf{S}_+$, the column space of common null covariance matrix of row vectors of $\mathbf{Q}$. Finally, the test statistic D^2 is the squared sum of Mahalanobis distances from the projected vectors to the origin, characterizing *differential departure from independence*:

$$D^2 = \sum_{m=1}^{(r-1)(s-1)} \left\| \mathbf{q}_m^\top \mathbf{S}_+ \Lambda_+^{-1/2} \right\|^2 \tag{6}$$

where $\mathbf{S}_+$ and Λ_+ are non-zero eigenvector and eigenvalue matrices of the covariance matrix of row vector $\mathbf{q}_m$ in $\mathbf{Q}$.

When the sample size is large, the amount added to each entry $1/(rs)$ will not influence the asymptotic distribution under the null hypothesis. Therefore, the compensated test statistic D^2 asymptotically follows the same a chi-squared distribution with $\nu = (K-1)(r-1)(s-1)$ degrees of freedom under the null hypothesis of row and column variables being independent as established for the original Sharma-Song test. The test can detect epistatic interactions either with or without marginal differences, but it rejects differential joint distribution only due to marginal differences.

The original test is effective in finding 2nd-order differential patterns in gene expression data across conditions (Sharma et al., 2020, 2021). When appropriate discretization of gene expression value is applied, it is less likely to obtain nearly empty rows or columns that violate the Cochran's condition of at least five expected count in each table entry. In GWAS, however, a low frequency in genotype or minor allele will not meet the Cochran's condition. This has a high impact on the test statistic of the original test. The compensated test improves over the original test by demoting patterns that appear 2nd-order differential caused by a violation. Figure 2 illustrates the improvement. Both tables violate the Cochran's condition. Patterns of both tables only differ in the genotype of aa-bb by one sample. The original test declares significant epistasis whereas the compensated test rejects epistasis.

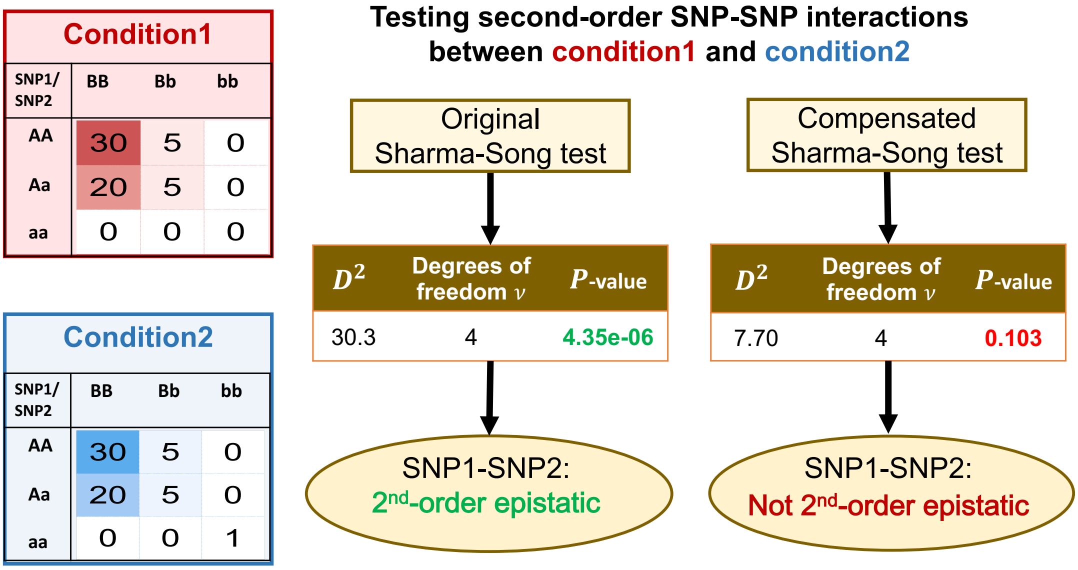
2.3 Simulating differential contingency tables

Now we present algorithms to simulate first-, second-, and full order differential patterns. They are implemented as the R function 'simulate_diff_tables' in the package 'DiffXTables' (version $\geq$ 0.1.0) from CRAN (Sharma and Song, 2021). Each type of differential patterns across K conditions is generated first by K population distributions that are required by the definition of the pattern type. Each population distribution is used as the probability parameters of a multinomial distribution to sample a table under each condition.

We first design Algorithm 1 SECOND-ORDER-ADJUST to generate 2nd-order adjustment to a 2×2 sub-matrix inside the population joint probability matrix. It imposes 2nd-order differences without changing marginal distributions.

2.3.1 Simulating first-order differential contingency tables

Algorithm 2 SIMULATE-1ST-ORDER-DIFFERENTIAL-TABLES generates K tables of given dimensions and a sample size such that the K tables are only first-order differential. The tables carry the same amount of joint probability deviation from the product of the marginal distributions, so that row and column variables are not necessarily independent. A positive integer parameter B controls the strength of 2nd-order component in the tables. A larger B allows more iterations to apply an overall stronger 2nd-order adjustment.



Condition1

SNP1/ SNP2	BB	Bb	bb
AA	30	5	0
Aa	20	5	0
aa	0	0	0

Condition2

SNP1/ SNP2	BB	Bb	bb
AA	30	5	0
Aa	20	5	0
aa	0	0	1

Testing second-order SNP-SNP interactions between condition1 and condition2

Original Sharma-Song test

D^2	Degrees of freedom ν	P-value
30.3	4	4.35e-06

SNP1-SNP2: 2nd-order epistatic

Compensated Sharma-Song test

D^2	Degrees of freedom ν	P-value
7.70	4	0.103

SNP1-SNP2: Not 2nd-order epistatic

Fig. 2 The compensated Sharma-Song test improves the original test. Genotype tables observed for a pair of SNPs under two conditions contain an almost identical pattern, differing only in genotype aa-bb with an expected occurrence of less than 5, violating the Cochran's condition. At significance level 0.05, the original test admits epistasis (P-value=4.3 × 10^{-6}), whereas the compensated test dismisses epistasis (P-value=0.103), overcoming the deficiency of the original test.

Algorithm 1 SECOND-ORDER-ADJUST($(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2), p_{\min}$)

```
$\Delta\mathbf{P} = [0]$
if no two adjacent entries are zero
    if all four entries are non-zero
        Select randomly an entry $c$ out of $(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2)$
    elseif there is only one zero entry
        Select that entry as $c$
    else // two zeros are diagonal to each other
        Select any of the diagonal entry as $c$
    Generate $\varepsilon$ randomly within the range $(0, p_{\min})$
    Add $\varepsilon$ to $c$: $\Delta\mathbf{P}[c] = +\varepsilon$
    Add $\varepsilon$ to the diagonal entry of $c$: $\Delta\mathbf{P}[\text{diag}(c)] = +\varepsilon$
    Subtract $\varepsilon$ from the horizontally adjacent entry of $c$: $\Delta\mathbf{P}[\text{h.adj}(c)] = -\varepsilon$
    Subtract $\varepsilon$ from the vertically adjacent entry of $c$: $\Delta\mathbf{P}[\text{v.adj}(c)] = -\varepsilon$
return $\Delta\mathbf{P}$
```

2.3.2 Simulating second-order differential contingency tables

Algorithm 3 SIMULATE-2ND-ORDER-DIFFERENTIAL-TABLES generates K tables of given dimensions and a sample size such that the K tables are only 2nd-order differential. The tables contain *different* amounts of joint probability deviation from the product of the marginal distributions. All tables share the same row and column marginal distributions. A positive integer parameter B controls the strength of 2nd-order component in the tables. A larger B allows more iterations to apply an overall stronger 2nd-order adjustment.

2.3.3 Simulating full-order differential contingency tables

Algorithm 4 SIMULATE-FULL-ORDER-DIFFERENTIAL-TABLES generates K tables of given dimensions and a sample size such that the K tables are both first- and 2nd-order differential. The tables have distinct row and column marginal distributions. The tables contain *different* amounts of joint probability deviation from the product of the marginal distributions. A positive integer parameter B controls the strength of 2nd-order component in the tables. A larger B allows more iterations to apply an overall stronger 2nd-order adjustment.

Algorithm 2 SIMULATE-1ST-ORDER-DIFFERENTIAL-TABLES(K, r, s, n, B)

```
Input: Number of tables $K$, table dimension $r \times s$ ($r$, $s \geq 2$), sample size $n \geq 4$,
number of iterations $B$
Output: $K$ first-order differential contingency tables $\mathbf{C}_1, \ldots, \mathbf{C}_K$
for $k = 1$ to $K$:
    Generate a random row marginal probability function $p_X(i)$, $i$=1 to $r$
    Generate a random column marginal probability function $p_Y(j)$, $j$=1 to $s$
    $\mathbf{P}^k_{ij} = p_X(i) \cdot p_Y(j)$ for $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, s\}$
repeat
    Generate random row indices $i_1 \neq i_2$ and column indices $j_1 \neq j_2$
    Define four entries indexed by $(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2)$
    Find $p_{\min} = \min_{k \in 1,\ldots,K}$ non-zero $\min\{\mathbf{P}^k_{i_1,j_1}, \mathbf{P}^k_{i_1,j_2}, \mathbf{P}^k_{i_2,j_1}, \mathbf{P}^k_{i_2,j_2}\}$
    $\Delta\mathbf{P}$ = SECOND-ORDER-ADJUST($(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2), p_{\min}$)
    $feasible$ = TRUE
    for $k = 1, \ldots, K$:
        $\mathbf{P}^{*k} = \mathbf{P}^k + \Delta\mathbf{P}$
        if some entries in $\mathbf{P}^{*k}$ are negative: $feasible$ = FALSE
    if $feasible$ == TRUE: for $k = 1, \ldots, K$: $\mathbf{P}^k = \mathbf{P}^{*k}$
until $B$ times
for $k = 1, \ldots, K$:
    Use $\mathbf{P}^k$ as multinomial probabilities to generate table $\mathbf{C}_k$ of sample size $n$
return $\mathbf{C}_1, \ldots, \mathbf{C}_K$
```

Algorithm 3 SIMULATE-2ND-ORDER-DIFFERENTIAL-TABLES(K, r, s, n, B)

```
Input: Number of tables $K$, table dimension $r \times s$ ($r$, $s \geq 2$), sample size $n \geq 4$,
number of iterations $B$
Output: $K$ second-order differential tables $\mathbf{C}_1, \ldots, \mathbf{C}_K$
Generate a random row marginal probability function $p_X(i)$, $i$=1 to $r$
Generate a random column marginal probability function $p_Y(j)$, $j$=1 to $s$
for $k = 1$ to $K$:
    $\mathbf{P}^k_{ij} = p_X(i) \cdot p_Y(j)$ for $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, s\}$
    repeat
        Generate random row indices $i_1 \neq i_2$ and column indices $j_1 \neq j_2$
        Define four entries indexed by $(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2)$
        $p_{\min}$ = the non-zero minimum from $\mathbf{P}^k_{i_1,j_1}, \mathbf{P}^k_{i_1,j_2}, \mathbf{P}^k_{i_2,j_1}, \mathbf{P}^k_{i_2,j_2}$
        $\mathbf{\Delta P}$ = SECOND-ORDER-ADJUST($(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2), p_{\min}$)
        $\mathbf{P}^k = \mathbf{P}^k + \mathbf{\Delta P}$
    until $B$ times
    Use $\mathbf{P}^k$ as multinomial probabilities to generate table $\mathbf{C}_k$ of sample size $n$
return $\mathbf{C}_1, \ldots, \mathbf{C}_K$
```

3 Results

3.1 Performance evaluation on synthetic data

3.1.1 Detecting epistasis simulated by GAMETES

To evaluate the effectiveness of Sharma-Song tests on detecting epistatic interactions, we first used a third-party software tool GAMETES 2.0 (Urbanowicz et al., 2012) to simulate epistatic data. GAMETES is a software package that generates pure, strictly biallelic and n-locus epistasis models with given heritability, minor allele frequency (MAF), and population prevalence. An epistasis model is represented by a penetrance table which contains the probability of an individual affected by the disease or trait at a particular genotype. It is denoted by $P(D|G_i)$, where D represents disease and G_i the genotype i. Three parameters, disease prevalence $P(D)$, genetic heritability h^2, and MAF, are used to control the penetrance value.

As the compensated Sharma-Song test is a non-parametric method without any model assumption, we utilized all parameter combinations of GAMETES without restricting ourselves to a fixed model.

Given a set of parameters, GAMATES can generate different penetrance tables with low to high detection difficulty with an option of Ease of Detection Measure (EDM). We performed two simulation studies with high and low MAF, respectively. In the first simulation, we generated 18 models from parameters h^2 = 0.01, 0.025, 0.05, 0.1, 0.2, 0.4 and MAF = 0.2, 0.3, 0.4. In the second simulation, we generated six models from parameters h^2 = 0.005, 0.02 and MAF =

Algorithm 4 SIMULATE-FULL-ORDER-DIFFERENTIAL-TABLES(K, r, s, n, B)

```
Input: Number of tables $K$, table dimension $r \times s$ ($r, s \geq 2$), sample size $n \geq 4$,
number of iterations $B$
Output: $K$ full-order differential tables $\mathbf{C}_1, \ldots, \mathbf{C}_K$
for $k = 1$ to $K$:
    Generate a random row marginal probability function $p_X(i)$, $i$=1 to $r$
    Generate a random column marginal probability function $p_Y(j)$, $j$=1 to $s$
    $\mathbf{P}^k_{ij} = p_X(i) \cdot p_Y(j)$ for $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, s\}$
    repeat
        Generate random row indices $i_1 \neq i_2$ and column indices $j_1 \neq j_2$
        Define four entries indexed by $(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2)$
        $p_{\min}$ = the non-zero minimum from $\mathbf{P}^k_{i_1,j_1}, \mathbf{P}^k_{i_1,j_2}, \mathbf{P}^k_{i_2,j_1}, \mathbf{P}^k_{i_2,j_2}$
        $\Delta\mathbf{P}$ = SECOND-ORDER-ADJUST($(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2), p_{\min}$)
        $\mathbf{P}^k = \mathbf{P}^k + \Delta\mathbf{P}$
    until $B$ times
    Use $\mathbf{P}^k$ as multinomial probabilities to generate table $\mathbf{C}_k$ of sample size $n$
return $\mathbf{C}_1, \ldots, \mathbf{C}_K$
```

0.05, 0.01, 0.1. For each parameter set, we utilized the penetrance table with difficult models of low EDM. A selected penetrance table was used to generate 200 datasets each with a pair of SNPs for 60 samples. We also generated a pair of unassociated SNPs with minimum MAF ranging from 0.05 to 0.1 with 60 samples. The associated and unassociated SNP pairs constitute the true and false pairs in the ground truth.

We compared the performance of the two Sharma-Song tests with the allele-based test (Purcell et al., 2007), BOOST (Wan et al., 2010) and the joint-effects test (Ueki and Cordell, 2012). The last three methods are implemented in the software package Plink (Purcell et al., 2007). The SNP pairs generated by GAMATES were converted into '.ped' files as input to Plink, whereas for the Sharma-Song tests, they were converted to contingency tables.

Figure 3a,b shows receiver operating characteristic (ROC) curves and precision-recall (PR) curves of all five methods at high MAF (0.2–0.4) at a wide range of heritability (0.01–0.4). Both Sharma-Song tests performed markedly better than other methods even at small sample sizes. Here the original Sharma-Song test does have a slight advantage over the compensated but their top ranked candidates are accurate (the very left parts of the curves). Figure 3c,d shows the ROC and PR curves of each method at low MAF (0.05, 0.01, 0.1), where variants can be marginally nonuniform to violate the Cochran's condition. The low heritability is a consequence of the low MAF. The compensated Sharma-Song test outperformed all other methods; it has improved over the original Sharma-Song test at lowered false positive rate (Fig 3c) and increased precision (Fig 3d) for the top ranked candidates. This result is of practically beneficial consequences, as a low MAF is often observed with diseases (Kido et al., 2018).

3.1.2 Detecting second- or full-order differential patterns against first-order difference

To determine if two variants are epistatic against having an independent marginal effect, we performed another simulation study to test second- or full-order differential patterns against first-order ones. Using Algorithms 2, 3, and 4 for simulating 1st-, 2nd-, and full-order differential patterns, we generated 500 pairs of 3×3 tables ($K = 2$) with sample size ranging from 50 to 300 at differentiality $B = 300$. The performance of all five methods is given in Fig. 4. Figure 4a,b is the ROC and PR curves for each method in recognizing 2nd- over 1st-order differential patterns. Both Sharma-Song tests evidently outperformed other methods. The lead of both Sharma-Song tests is even more remarkable in Figure 4c,d for detecting full-order over 1st-order differential patterns. This setting represents a highly expected scenario in real applications. The second-order advantage is a distinct feature of the Sharma-Song tests, not previously implemented in any other tests known to us.

3.2 SNP-SNP epistasis associated with chicken fat content

To illustrate the diversity of epistasis patterns selected by different methods, we evaluated pairwise patterns among 48,034 SNPs surveyed in 475 male chickens of contrasting abdominal fat content. The chicken GWAS dataset (Li et al., 2013) was collected from the Northeast Agricultural University broiler line NEAUHLF divergently selected for abdominal fat content (Guo et al., 2011). Out of the 475 chickens, 203 are in the lean line and 272 in the fat line, both from the 11th generation population of NEAUHLF. Samples with 5% or more loci missing genotype information were removed; loci with less than 5% MAF were also filtered out from the published data. Each genotype was represented as one of AA, AB

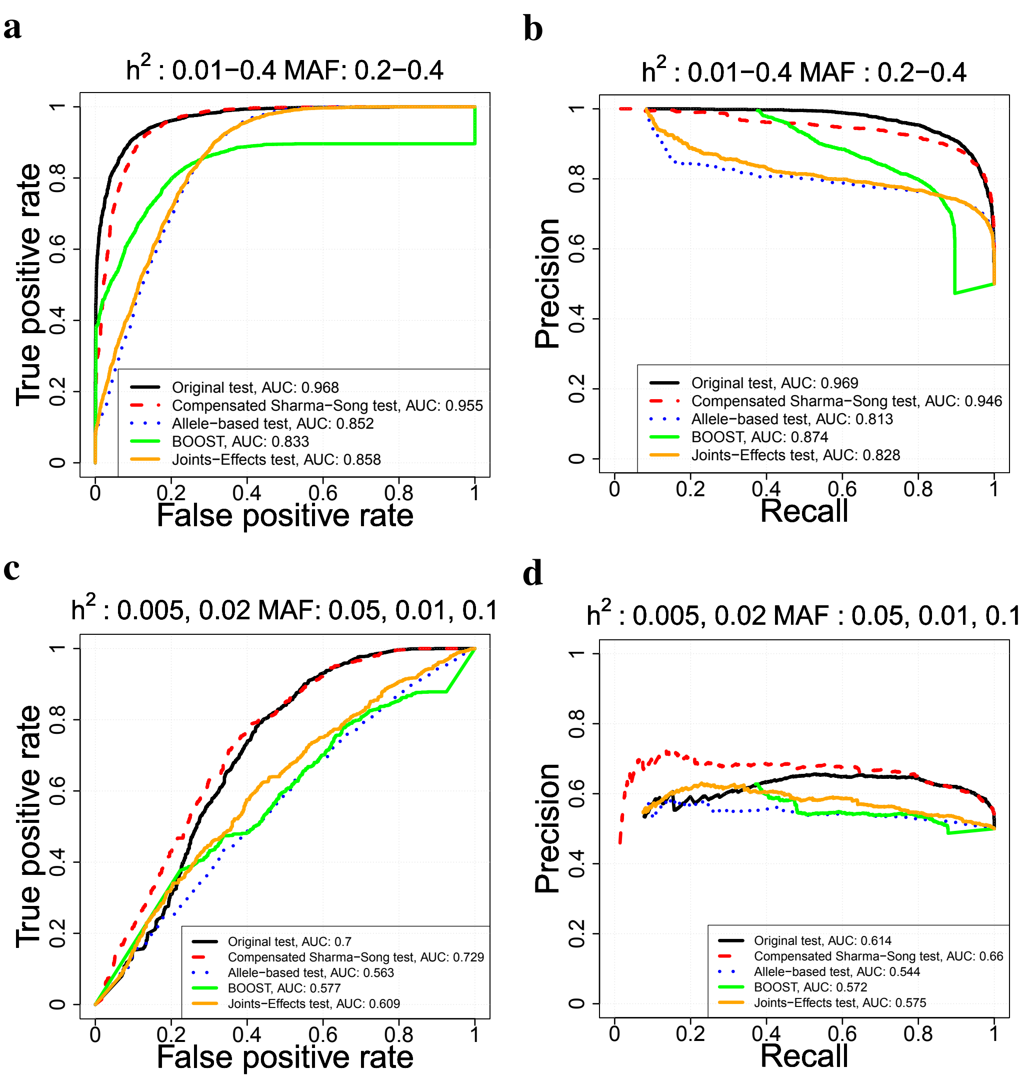


Fig. 3 Performance on detecting joint from marginal effects using GAMETES epistasis models by all five methods. (a,b) The ROC and PR curves of each method at relatively high minor allele frequencies. **(c,d)** The ROC and PR curves of each method at relatively low minor allele frequencies.

and *BB* allele combinations (Zhang et al., 2012). The *B* allele frequency of each locus is used to determine the minor allele in the population.

3.2.1 Genetic epistatic patterns associated with chicken abdominal fat

To identify complex epistatic patterns at the second order, we evaluated $\binom{48,034}{2}$ SNP-SNP patterns using the compensated Sharma-Song test (Sharma et al., 2021; Sharma and Song, 2021), allele-based test (Purcell et al., 2007), BOOST (Wan et al., 2010) and the joint-effects test (Ueki and Cordell, 2012).

For the compensated Sharma-Song test, after evaluating all SNP pairs, we obtained 98,823,786 significant epistatic interactions at Benjamini-Hochberg (Benjamini and Hochberg, 1995) adjusted P-value ≤ 0.05. To utilize the methods implemented in Plink software, we first converted genotype information of all the SNPs in '.ped' and '.map' format as input to Plink. We also created a text file to provide the phenotype of each chicken subject. We called the fast-epistasis option for all three methods with P-value ≤ 0.05. For the joint-effects test, we set the minimum sample per cell requirement to 0. While calling the allele-based test, BOOST and the joint-effects test, Plink removed 136 monomorphic SNP markers. Finally, 67,824,629 significant interactions are obtained by the allele-based test, 20,886,382 by BOOST, and 50,888,329 by the joint-effects test, all at Benjamini-Hochberg (Benjamini and Hochberg, 1995) adjusted P-value$\leq$0.05. Genotype and phenotype data of the top three ranked patterns of each method are visualized in Fig. 5. The compensated Sharma-Song test detected three epistatic interactions between SNPs of same chromosome (Fig. 5a); these SNPs, not in linkage disequilibrium according to Plink, are additive by additive epistatic patterns that are strong and clean. BOOST picked

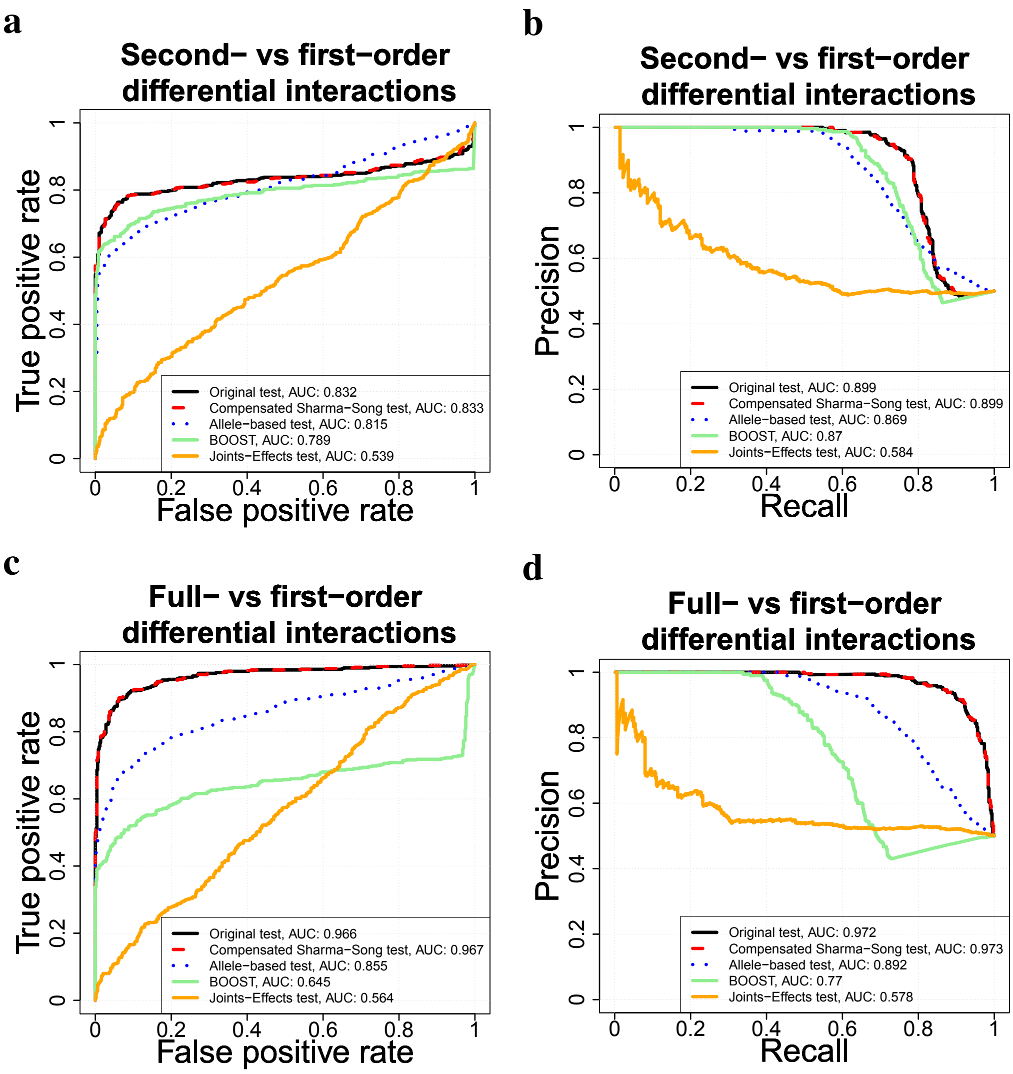


Fig. 4 Performance on detecting second-order or full-order against first-order patterns. (a,b) The ROC and PR curves of five methods for differentiating second- versus first-order patterns. **(c,d)** The ROC and PR curves of each method for telling apart full-order from first-order patterns.

patterns where the two SNPs are independent and one of them being additive but not epistatic (Fig. 5b). The allele-based test promoted weak epistatic patterns which violated the Cochran's condition (Fig. 5c). The joint-effects test revealed patterns that are additive by additive epistatic (Fig. 5d), but not as strong or clean as those returned by the compensated Sharma-Song test.

A genomic map of the detected SNP-SNP epistasis associated with chicken fat detected by the compensated Sharma-Song test is visualized by Fig. 6a, covering chromosomes 1 to 28 and Z of chicken. At a P-value threshold of 1×10^{-20}, a total of 270,828 significant SNP-SNP interactions are found; SNPs of 3,332 pairs are from different chromosomes and 267,496 of them in the same one. We found 682 SNP-SNP interactions out of 270,828 interactions that are in linkage disequilibrium with $R^2 \geq 0.9$.

3.2.2 Pathways enriched in epistatic interactions

We report pathways that are highly enriched in genes involved in top SNP-SNP epistatic interactions. We selected top epistatic interactions with Benjamini-Hochberg (Benjamini and Hochberg, 1995) adjusted P-value $\leq 1 \times 10^{-20}$ and mapped the SNPs to 9,760 unique enclosing genes, using a genome annotation file of the chicken *Gallus gallus* from the UCSC Genome Browser. Using the 'KEGGREST' package (Tenenbaum and Maintainer, 2021), we collected all chicken pathways and genes. We performed 'SIGORA' pathway analysis (Foroushani et al., 2013) and obtained 42 significant pathways at Bonferroni (Bonferroni, 1935) adjusted P-value ≤ 0.05. Table 1 lists the top five enriched pathways with their description, adjusted P-value and number of successes: the MAPK signaling pathway (gga04010), endocytosis

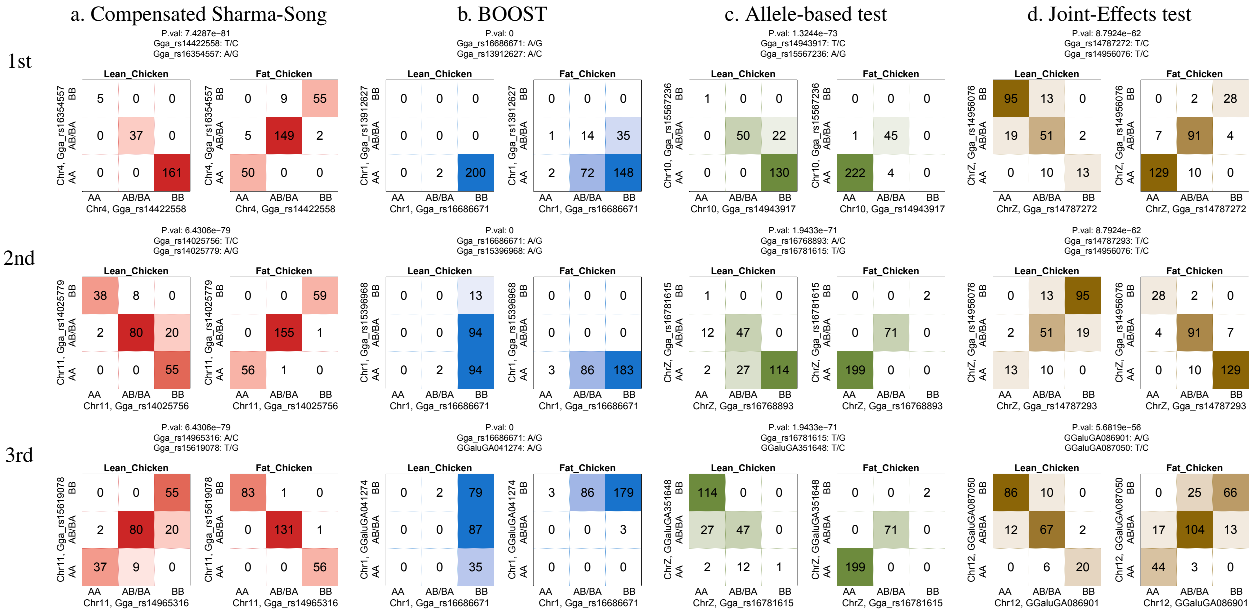


Fig. 5 Top three SNP-SNP epistasis patterns associated with chicken abdominal fat content inferred by four methods. (a) The compensated Sharma-Song test. **(b)** BOOST. **(c)** The allele-based test. **(d)** The joint-effects test. Each 3×3 contingency table tabulates the genotype frequency in lean or fat chicken. The color intensity of each table entry is proportional to the count in that entry.

Table 1 Top five KEGG pathways enriched in significant epistatic interactions associated with chicken abdominal fat.

Pathway ID	Description	Adjusted *P*	Successes	#Genes
gga04010	MAPK signaling pathway	0.0e+00	170	260
gga04144	Endocytosis	1.4e-296	146	232
gga04510	Focal adhesion	2.5e-226	138	191
gga04310	Wnt signaling pathway	5.6e-147	101	147
gga04512	ECM-receptor interaction	7.8e-115	63	84

(gga04144), focal adhesion (gga04510), Wnt signaling pathway (gga04310), and ECM-receptor interaction (gga04512). We examined common genes between these five pathways and observed 11 shared genes between the first two pathways. However, no shared gene was found among all five pathways. Most pathways are directly or indirectly associated with obesity. The MAPK signaling pathway (gga04010) is effected in monogenic obesity, as mutation in the melanocortin-4 receptor (*MC4R*) gene, leading to obesity in humans, causes defects in MAPK signaling (He and Tao, 2014). One enriched pathway of endocytosis regulates various metabolic activities such as glucose and lipid metabolism in mice and humans (Gilleron et al., 2019). Mutation in *MC4R* in human can impact receptor homodimerization, endocytosis, and trafficking which in turn affects obesity in human (Brouwers et al., 2021). Adipose tissue is a connective tissue containing lipid cells known as adipocytes. One study showed the importance of focal adhesion kinase (FAK) in adipocyte survival and maintaining insulin sensitivity in expansion of adipose tissue in obese mice and humans (Luk et al., 2017). One study on chicken found that fat mass and obesity associated genes enhance the differentiation of myoblasts by regulating eight genes in the focal adhesion pathway (Huang et al., 2020). The ECM receptor pathway of adipocytes and other cells in adipose tissue are critical in regulating the intracellular signaling of angiogenesis, adipocyte death and the infiltration of inflammatory cells, which culminate in insulin resistance (Lin et al., 2016). The Wnt signaling pathway regulates early developmental processes via mediating cell differentiation, cell proliferation (Alonso and Fuchs, 2003). Studies have linked variation in the Wnt signaling pathway with obesity during the high-caloric diet in the depot-specific role of adipose tissue (Loh et al., 2015; Chen and Wang, 2018).

3.2.3 A network of top epistatic SNP-SNP interactions

We extracted 15 hub SNPs with highest degrees from the top 270,828 significant SNP-SNP patterns. We sampled 20 patterns for top 10 hub SNPs to create a subnetwork of 200 interactions (Fig. 6b). Table 2 shows the top 15 hub SNPs with chromosome number, position, allele, gene they belong to, and degree (the number of times they are epistatic with another SNP). The very first SNP (GGaluGA243716) on chromosome 4 does not belong to any gene; it is most likely in

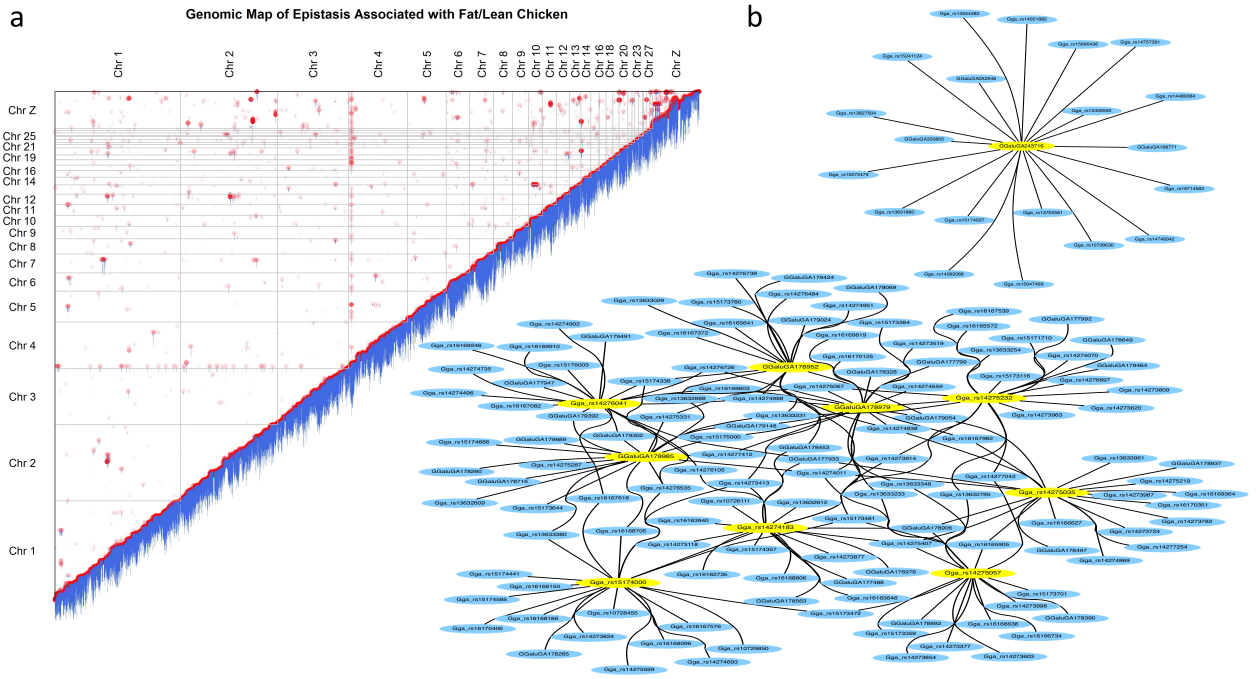


Fig. 6 Top SNP-SNP epistatic interactions detected by the compensated Sharma-Song test associated with chicken abdominal fat. (a) A genomic map of most significant 270,828 epistatic SNP-SNP interactions with adjusted P-value $\leq 1 \times 10^{-20}$. Each pair is shown as a red dot with a blue tail. The intensity of the dot and the length of the tail are proportional to the negative log of P-value. Although most SNP-SNP patterns are found within the same chromosome, many patterns are across chromosomes. Of note, a SNP GGaluGA243716 in a locus at the beginning of chromosome 4 is interacting with most other chromosomes. **(b)** A subnetwork formed by 200 significant epistatic interactions. Each node is a SNP. Each edge is a statistically significant epistatic pattern involving its two SNP nodes. Hub SNPs are highlighted by nodes in a yellow background.

Table 2 Fifteen SNP hubs of highest degrees in a network formed by 270,828 SNP-SNP pairs of most significant epistatic patterns associated with chicken abdominal fat.

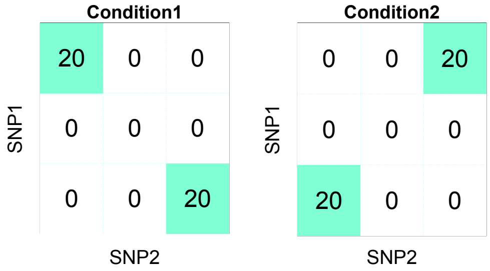
SNP	Chromosome:position	Allele	Gene	Degree
GGaluGA243716	4: 4724264	[A/G]	-	152
Gga_rs15174000	20: 6468949	[A/C]	-	136
Gga_rs14275232	20: 7427409	[T/C]	*CDH4*	134
Gga_rs14275035	20: 7225147	[A/G]	-	130
Gga_rs14275057	20: 7236679	[A/G]	*CDH4*	129
GGaluGA178952	20: 8147159	[A/G]	-	126
GGaluGA178979	20: 8198520	[A/G]	*SLCO4A1*	126
GGaluGA178985	20: 8206011	[A/G]	*SLCO4A1*	126
Gga_rs14276041	20: 8074520	[A/G]	-	125
Gga_rs14274183	20: 6431557	[A/C]	-	124
Gga_rs13633380	20: 7449839	[A/G]	*CDH4*	122
Gga_rs15174939	20: 7434302	[T/C]	*CDH4*	122
GGaluGA178563	20: 7570711	[A/G]	*CDH4*	116
GGaluGA178919	20: 8102870	[T/C]	-	116
Gga_rs14025978	11: 14463528	[T/C]	*WWOX*	114

an intergenic region. Its interaction with other SNPs is highly visible in Fig. 6a. This SNP is of high potential for a role in chicken fat content.

It is intriguing that most of the top hub SNPs lie in chromosome 20. A hub SNP Gga_rs14025956 on chromosome 11 forming 114 epistatic interactions is aligned to the *WWOX* gene, a tumor suppressor gene. A recent mouse study found that the many *WWOX* variants in the skeletal muscle result in weight gain, insulin resistance and glucose intolerance (Abu-Remaileh et al., 2019).

4 Discussion

A critical advantage of the compensated Sharma-Song test is its robustness to the Cochran's condition (having an expected count of at least 5 in each table entry). Other methods must deal with this condition explicitly by skipping tables violating the condition. However, a pair of tables not meeting the Cochran's condition may still be strongly epistatic. One such example is shown in Figure 7. The overall pattern is ideally additive by additive epistatic (Fig. 1c). Excluding such tables can degrade the power of an epistasis test. In the compensated test, the addition of the small value $1/(rs)$ (1/9 for genotype epistasis) does not alter data in any substantial way, but it aggressively demotes uninteresting patterns which would be judged epistatic by other tests only due to violation of the Cochran's condition (Fig. 2). Therefore, the compensated test being robust to the Cochran's condition makes it highly practical for epistasis analysis.



Condition1

SNP1 \ SNP2			
	20	0	0
	0	0	0
	0	0	20

Condition2

SNP1 \ SNP2			
	0	0	20
	0	0	0
	20	0	0

Fig. 7 Perfect epistasis that violates the Cochran's condition. Such patterns could have been filtered out by a standard preprocessing pipeline. The compensated Sharma-Song test declares the pair as second-order differential (P-value 7.5×10^{-16}), exemplifying ideal additive by additive epistasis.

To apply the compensated Sharma-Song test on quantitative traits, continuous phenotypic values can be discretized either independently by methods such as 'Ckmeans.1d.dp' (Song and Zhong, 2020; Song et al., 2022) or jointly by methods such as 'GridOnClusters' (Wang et al., 2020, 2022).

5 Conclusions

The compensated Sharma-Song test is unique in detecting differential departure from independence, resistant to joint distributional differences caused only by marginal changes. It penalizes patterns which would be false positive epistasis by other tests due to the unmet Cochran's condition, hence eliminating the need for a preprocessing step to filter out those patterns. What is less obvious is its capability to recognize patterns that are epistatic but not satisfying the Cochran's condition. After generating a map of SNP-SNP epistasis associated with chicken abdominal fat content, the test revealed uniquely strong additive-by-additive epistasis. We found most hub SNPs on chromosome 20 and some in intergenic regions. Despite the many existing methods for epistasis analysis, the compensated Sharma-Song test is a practical method for uncovering complex genotype-phenotype effects in genome-wide association studies.

6 Acknowledgments

This reported work was partially supported by National Science Foundation grant 1661331 and USDA grant 2016-51181-25408.

Data and code availability

The compensated Sharma-Song test and the differential table simulation algorithms are implemented in the open-source R package 'DiffXTables' (Sharma and Song, 2021) freely available from https://cran.r-project.org/package=DiffXTables. Data and other source code are available at Code Ocean doi: 10.24433/CO.7661508.v1.

References

Abu-Remaileh M, Abu-Remaileh M, Akkawi R, Knani I, Udi S, Pacold ME, Tam J, Aqeilan RI (2019) WWOX somatic ablation in skeletal muscles alters glucose metabolism. Molecular Metabolism 22:132–140, DOI 10.1016/j.molmet.2019.01.010

Alonso L, Fuchs E (2003) Stem cells in the skin: waste not, Wnt not. Genes & Development 17(10):1189–1200, DOI 10.1101/gad.1086903

Andreasen NC, Wilcox MA, Ho BC, Epping E, Ziebell S, Zeien E, Weiss B, Wassink T (2012) Statistical epistasis and progressive brain change in schizophrenia: an approach for examining the relationships between multiple genes. Molecular Psychiatry 17(11):1093–1102, DOI 10.1038/mp.2011.108

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300, DOI 10.2307/2346101

Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste. In: Studi in Onore del Professore Salvatore Ortu Carboni, Rome: Italy, pp 13–60

Brouwers B, de Oliveira EM, Marti-Solano M, Monteiro FB, Laurin SA, Keogh JM, Henning E, Bounds R, Daly CA, Houston S, Ayinampudi V, Wasiluk N, Clarke D, Plouffe B, Bouvier M, Babu MM, Farooqi IS, Mokrosiński J (2021) Human MC4R variants affect endocytosis, trafficking and dimerization revealing multiple cellular mechanisms involved in weight regulation. Cell Reports 34(12):108862, DOI 10.1016/j.celrep.2021.108862

Chen N, Wang J (2018) Wnt/β-catenin signaling and obesity. Frontiers in Physiology 9:792, DOI 10.3389/fphys.2018.00792

Cho Y, Ritchie M, Moore J, Park J, Lee KU, Shin H, Lee H, Park K (2004) Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia 47(3):549–554, DOI 10.1007/s00125-003-1321-3

Cochran WG (1952) The χ^2 test of goodness of fit. The Annals of Mathematical Statistics 23(3):315–345, DOI 10.1214/aoms/1177729380

Cochran WG (1954) Some methods for strengthening the common χ^2 tests. Biometrics 10(4):417–451, DOI 10.2307/3001616

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics 11(20):2463–2468, DOI 10.1093/hmg/11.20.2463

Domingo J, Baeza-Centurion P, Lehner B (2019) The causes and consequences of genetic interactions (epistasis). Annu Rev Genomics Hum Genet 20:433–460, DOI 10.1146/annurev-genom-083118-014857

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96(456):1348–1360, DOI 10.1198/016214501753382273

Foroushani AB, Brinkman FS, Lynn DJ (2013) Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures. PeerJ 1:e229, DOI 10.7717/peerj.229

Gilleron J, Gerdes JM, Zeigerer A (2019) Metabolic regulation through the endosomal system. Traffic 20(8):552–570, DOI 10.1111/tra.12670

Guo L, Sun B, Shang Z, Leng L, Wang Y, Wang N, Li H (2011) Comparison of adipose tissue cellularity in chicken lines divergently selected for fatness. Poultry Science 90(9):2024–2034, DOI 10.3382/ps.2010-00863

He S, Tao YX (2014) Defect in MAPK signaling as a cause for monogenic obesity caused by inactivating mutations in the melanocortin-4 receptor gene. International Journal of Biological Sciences 10(10):1128, DOI 10.7150/ijbs.10359

Howard TD, Koppelman GH, Xu J, Zheng SL, Postma DS, Meyers DA, Bleecker ER (2002) Gene-gene interaction in asthma: IL4RA and IL13 in a dutch population with asthma. The American Journal of Human Genetics 70(1):230–236, DOI 10.1086/338242

Huang H, Liu L, Li C, Liang Z, Huang Z, Wang Q, Li S, Zhao Z (2020) Fat mass-and obesity-associated (FTO) gene promoted myoblast differentiation through the focal adhesion pathway in chicken. 3 Biotech 10(9):1–10, DOI 10.1007/s13205-020-02386-z

Jing PJ, Shen HB (2015) MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. Bioinformatics 31(5):634–641, DOI 10.1093/bioinformatics/btu702

Kempthorne O (1957) An Introduction to Genetic Statistics. Wiley publications in statistics, Wiley, New York

Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, Chen R, Sirota M, Kodama K, Hadley D, et al. (2018) Are minor alleles more likely to be risk alleles? BMC Medical Genomics 11(1):1–11, DOI 10.1186/s12920-018-0322-5

Li F, Hu G, Zhang H, Wang S, Wang Z, Li H (2013) Epistatic effects on abdominal fat content in chickens: results from a genome-wide SNP-SNP interaction analysis. PLoS One 8(12):e81520, DOI 10.1371/journal.pone.0081520

Lin D, Chun TH, Kang L (2016) Adipose extracellular matrix remodelling in obesity and insulin resistance. Biochemical Pharmacology 119:8–16, DOI 10.1016/j.bcp.2016.05.005

Loh NY, Neville MJ, Marinou K, Hardcastle SA, Fielding BA, Duncan EL, McCarthy MI, Tobias JH, Gregson CL, Karpe F, et al. (2015) LRP5 regulates human body fat distribution by modulating adipose progenitor biology in a dose-and depot-specific fashion. Cell Metabolism 21(2):262–273, DOI 10.1016/j.cmet.2015.01.009

Luk CT, Shi SY, Cai EP, Sivasubramaniyam T, Krishnamurthy M, Brunt JJ, Schroer SA, Winer DA, Woo M (2017) FAK signalling controls insulin sensitivity through regulation of adipocyte survival. Nature Communications 8(1):1–13, DOI 10.1038/ncomms14360

Ma L, Runesha HB, Dvorkin D, Garbe JR, Da Y (2008) Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. BMC Bioinformatics 9(1):1–9, DOI 10.1186/1471-2105-9-315

Niel C, Sinoquet C, Dina C, Rocheleau G (2015) A survey about methods dedicated to epistasis detection. Frontiers in Genetics 6:285, DOI 10.3389/fgene.2015.00285

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81(3):559–575, DOI 10.1086/519795

Shang J, Zhang J, Sun Y, Liu D, Ye D, Yin Y (2011) Performance analysis of novel methods for detecting epistasis. BMC Bioinformatics 12(1):1–17, DOI 10.1186/1471-2105-12-475

Sharma R, Song M (2021) 'DiffXTables': Pattern Analysis Across Contingency Tables. URL `https://CRAN.R-project.org/package=DiffXTables`, R package version 0.1.3

Sharma R, Luo X, Kumar S, Song M (2020) Three co-expression pattern types across microbial transcriptional networks of plankton in two oceanic waters. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Association for Computing Machinery, New York, NY, USA, BCB '20, DOI 10.1145/3388440.3412485

Sharma R, Kumar S, Song M (2021) Fundamental gene network rewiring at the second order within and across mammalian systems. Bioinformatics 37(19):3293–3301, DOI 10.1093/bioinformatics/btab240

Song J, Zhong H, Wang H (2022) 'Ckmeans.1d.dp': Optimal, Fast, and Reproducible Univariate Clustering. URL `https://CRAN.R-project.org/package=Ckmeans.1d.dp`, R package version 4.3.4

Song M, Zhong H (2020) Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers. Bioinformatics 36(20):5027–5036, DOI 10.1093/bioinformatics/btaa613

Tenenbaum D, Maintainer BP (2021) KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.32.0

Tuo S, Zhang J, Yuan X, He Z, Liu Y, Liu Z (2017) Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations. Scientific Reports 7(1):1–18, DOI 10.1038/s41598-017-11064-9

Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genet 8(4):e1002625, DOI 10.1371/journal.pgen.1002625

Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Mining 5(1):1–14, DOI 10.1186/1756-0381-5-16

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics 87(3):325–340, DOI 10.1016/j.ajhg.2010.07.021

Wang J, Kumar S, Song M (2020) Joint grid discretization for biological pattern discovery. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '20, DOI 10.1145/3388440.3412415

Wang J, Kumar S, Song J (2022) 'GridOnClusters': Cluster-Preserving Multivariate Joint Grid Discretization. URL `https://CRAN.R-project.org/package=GridOnClusters`, R package version 0.1.0

Winham S, Wang C, Motsinger-Reif AA (2011) A comparison of multifactor dimensionality reduction and L1-penalized regression to identify gene-gene interactions in genetic association studies. Statistical Applications in Genetics and Molecular Biology 10(1), DOI 10.2202/1544-6115.1613

Xie M, Li J, Jiang T (2012) Detecting genome-wide epistases based on the clustering of relatively frequent items. Bioinformatics 28(1):5–12, DOI 10.1093/bioinformatics/btr603

Yang C, He Z, Wan X, Yang Q, Xue H, Yu W (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics 25(4):504–511, DOI 10.1093/bioinformatics/btn652

Zhang H, Wang SZ, Wang ZP, Da Y, Wang N, Hu XX, Zhang YD, Wang YX, Leng L, Tang ZQ, et al. (2012) A genome-wide scan of selective sweeps in two broiler chicken lines divergently selected for abdominal fat content. BMC Genomics 13(1):1–16, DOI 10.1186/1471-2164-13-704

Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nature Genetics 39(9):1167–1173, DOI 10.1038/ng2110